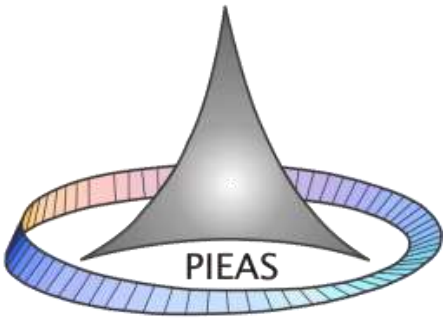
Intelligent Decision Making Ensemble Classification System for Breast Cancer Prediction



Safdar Ali

2015

Pakistan Institute of Engineering and Applied Sciences
Nilore, Islamabad, Pakistan

This page intentionally left blank.

Thesis SubmissionApproval

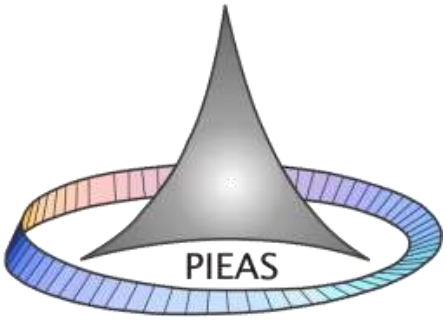
This is to certify that the work contained in this thesis entitled **Intelligent Decision Making Ensemble Classification System for Breast Cancer Prediction**, was carried out by **Safdar Ali**, and in my opinion, it is fully adequate, in scope and quality, for the degree of **Ph.D**. Furthermore, it is hereby approved for submission for review and thesis defense.

Supervisor: _____
Name: **Dr. Abdul Majid**
Date: 27 July, 2015
Place: PIEAS, Islamabad.

Co-Supervisor: _____
Name: **Dr. Asifullah Khan**
Date: 27 July, 2015
Place: PIEAS, Islamabad.

Head DCIS: _____
Name: **Dr. Javid Khurshid**
Date: 27 July, 2015
Place: PIEAS, Islamabad.

Intelligent Decision Making Ensemble Classification System for Breast Cancer Prediction



Safdar Ali

Submitted in partial fulfillment of the requirements
for the degree of Ph.D.
27 July 2015

Department of Computer and Information Sciences
Pakistan Institute of Engineering and Applied Sciences
Nilore, Islamabad, Pakistan

Dedications

Dedicated to my parents, wife, and daughter

Acknowledgements

Firstly, I would like to thank God Almighty for the blessings granted to me during this research work. He bestowed upon me strength, determination, and knowledge to complete my PhD research work.

I would like to convey my heartedly thanks to my supervisor, Dr. Abdul Majid, whose sincere support, encouragement and guidance, helped me in completion of my PhD research. He broadened my viewpoint in several aspects by sharing his knowledge with me. Additionally, his close assistance and encouraging behavior in tough times are real inspiration to reach the milestone timely. Moreover, I thank Dr. Asifullah Khan, DCIS (Co-supervisor) for productive cooperation, support, and valuable comments during this research. I would like to express my appreciation and heartfelt gratitude to Ex-Dean Research, PIEAS, Dr. Khalid Jamil for his motivation and moral boost to complete my PhD work. I am thankful to all teachers of the Department of DCIS and specially, Dr. Abdul Jalil and Dr. Mutawarra Hussain for their productive cooperation, support, and valuable comments during this research.

I am also thankful to colleagues at Directorate General of National Repository (DGNR) for their moral support, especially Mr. Masood Akhtar, who always encouraged me for higher studies. It is my great pleasure to show indebtedness to my friends like Professor Late Muhmmad Younas Awan, Professor Shahzad A. Malik, Awais Baig, Khaliq Awan, and PhD scholars of DCIS for their help during the course of this work. I am also thankful to all former students and colleagues (especially Dr. Aksam Iftikhar, Dr. Saima Rathore, Dr. Muhammad Tahir, Dr. khurram Jawad, Dr. Mehdi Hassan, Adnan, Jibran and Ahmad Ali) for their generous cooperation during PhD research work at PIEAS.

None of this would have been achievable without the support of my family. I would like to express my heartfelt gratitude to my loving parents, wife, and daughter whose heartfelt prayers and support made this work a success. Gratitude is also extended to my brother, sisters and in-laws for caring about my success and encouraging me in every possible way. Finishing this work also reminds me of my

late grandmother a lot, who would have loved seeing her grandson a successful person.

Finally, I would like to thank the Pakistan Atomic Energy Commission (PAEC), for financial support provided. I am also extremely thankful to Chairman, Member (Fuel Cycle), and Member (Materials), PAEC who allowed me to complete research work at PIEAS. Special thanks to Mr. Manzur Hussain, Director General, DGNR, for his infallible motivation and his support in allowing me to complete my dissertation.

(**Safdar Ali**)
PIEAS, Islamabad

Declaration of Originality

I hereby declare that the work contained in this thesis and the intellectual content of this thesis are the product of my own work. This thesis has not been previously published in any form nor does it contain any verbatim of the published resources which could be treated as infringement of the international copyright law. I also declare that I do understand the terms ‘copyright’ and ‘plagiarism,’ and that in case of any copyright violation or plagiarism found in this work, I will be held fully responsible of the consequences of any such violation.

_____

(**Safdar Ali**)
27 July, 2015
PIEAS, Islamabad.

Copyrights Statement

Table of Contents

List of Figures

List of Tables

Abstract

Breast cancer is a complex and heterogeneous disease which seriously impacts women's health. The diagnostic of breast cancer is an intricate process. Therefore, an accurate and reliable prediction system for breast cancer is indispensable to avoid misleading results.In this regard, improved decision support systems are essential for breast cancer prediction. Consequently, this thesis focuses on the development of intelligent decision making systems using ensemble classification for the early prediction of breast cancer.

Proteins of a breast tissue generally reflect the initial changes caused by successive genetic mutations, which may lead to cancer. In this research, such changes in protein sequences are exploited for the early diagnosis of breast cancer. It is found that substantial variation of Proline, Serine, Tyrosine, Cysteine, Arginine, and Asparagine amino acid molecules in cancerous proteins offer high discrimination for cancer diagnostic. Molecular descriptors derived from physicochemical properties of amino acids are used to transform primary protein sequences into feature spaces of amino acid composition (AAC), split amino acid composition (SAAC), pseudo amino acid composition-series (PseAAC-S), and pseudo amino acid composition-parallel (PseAAC-P).

The research work in this thesis is divided in two phases. In the first phase, the basic framework is established to handle imbalanced dataset in order to enhance true prediction performance. In this phase, conventional individual learning algorithms are employed to develop different prediction systems. Firstly, in conjunction with oversampling based Mega-Trend-Diffusion (MTD) technique, individual prediction systems are developed. Secondly, homogeneous ensemble systems "CanPro-IDSS" and "Can-CSCGnB" are developed using MTD and cost-sensitive classifier (CSC) techniques, respectively. It is found that assimilation of MTD technique for the CanPro-IDSS system is superior than CSC based technique to handle imbalanced dataset of protein sequences. In this connection, a web based *CanPro-IDSS* cancer

prediction system is also developed. Lastly, a novel heterogeneous ensemble system called “IDMS-HBC” is developed for breast cancer detection.

The second phase of this research focuses on the exploitation of variation of amino acid molecules in cancerous protein sequences using physicochemical properties. In this phase, unlike traditional ensemble prediction approaches, the proposed “IDM-PhyChm-Ens” ensemble system is developed by combining the decision spaces of a specific classifier trained on different feature spaces. This intelligent ensemble system is constructed using diverse learning algorithms of Random Forest (RF), Support Vector Machines, K-Nearest Neighbor, and Naïve Bayes (NB). It is observed that the combined spaces of SAAC+PseAAC-S and AAC+SAAC possess the best discrimination using ensemble-RF and ensemble-NB. Lastly, a novel classifier stacking based evolutionary ensemble system “Can-Evo-Ens” is also developed, whereby Genetic programming is used as the ensemble method. This study revealed that PseAAC-S feature space carries better discrimination power compared to AAC, SAAC, and PseAAC-P based feature extraction strategies.

Intensive experiments are performed to evaluate the performance of the proposed intelligent decision making systems for cancer/non-cancer and breast/non-breast cancer datasets. The proposed approaches have demonstrated improvement over previous state-of-the-art approaches. The proposed systems may be useful for academia, practitioners, and clinicians for the early diagnosis of breast cancer using protein sequences. Finally, it is expected that the findings of this research would have positive impact on diagnosis, prevention, treatment, and management of cancer.

List of Publications

Journal Articles:

- Abdul Majid, **Safdar Ali**, Mubashar Iqbql, Nabeela Kausar, "Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines," *Computer Methods and Programs in Biomedicine*, vol. 113, pp. 792-808, 2014. Impact Factor 1.555.
- **Safdar Ali**, Abdul Majid, Asifullah Khan, "IDM-PhyChm-Ens: Intelligent decision-making ensemble methodology for classification of human breast cancer using physicochemical properties of amino acids," *Amino Acids*, vol 46 (4), pp.977-993, 2014. Impact Factor 3.914.
- Abdul Majid and **Safdar Ali**, "HBC-Evo: Predicting human breast cancer by exploiting amino acid sequence based feature spaces and evolutionary ensemble," *Amino Acid*, vol. 47 (1), pp.217-221, 2015. Impact Factor 3.914.
- **Safdar Ali**, Abdul Majid, "Can-Evo-Ens: Classifier stacking based evolutionary ensemble system for prediction of human breast cancer using amino acid sequences," *Journal of Biomedical Informatics*, vol. 54, pp. 256-69, 2015. Impact Factor 2.482.
- **Safdar Ali** and Abdul Majid, "IDMS-HBC: Intelligent decision making ensemble system for prediction of human breast cancer form imbalanced data," *Medical & Biological Engineering & Computing*, under review, 2015.
- **Safdar Ali** and Abdul Majid, "CanPro-IDSS: Intelligent decision making classification models for breast cancer detection by exploiting discriminant information of cancer protein molecules and imbalanced data," *Journal of chemical information and modeling*, under review, 2015.
- Abdul Majid, **Safdar Ali**, Syed Gibran Javed, and Mohsin Sattar, "Can-CSCGnB: Developing Cost-Sensitive gentle ensemble system for breast cancer classification using protein amino acids and imbalanced data," *Journal of Protein & Peptide Letters*, under review, 2015.

Co-Supervision of M. Phil/MS Thesis:

- Sardar Ali Shah, session 2012-14, thesis title: "Development of Fuzzy Inference System in the Application of Safety Assessment of Low Level Radioactive Waste Repository," *M.Phil physics, Department of Physics and Applied Mathematics (DPAM)*, PIEAS.
- Fahad Ahmed, session 2012-14, thesis title: Segmentation of Fractures in Rock Images for Radioactive Waste Repository, *MS system engineering, Department of Electrical Engineering (DEE)*, PIEAS.

List of Abbreviations and Symbols

Symbol / notation	Description
AAC	Amino Acid Composition
Acc	Accuracy
AUCH	Area Under Convex Hull
AUC-ROC	Area Under Curves of Receiver Operating Characteristic
B/NBC	Breast/Non-Breast Cancer
C/NC	Cancer/Non-Cancer
CC/NCC	Colon/Non-Colon Cancer
CSL/CSC	Cost-Sensitive Learning/Classifier
dTIe	Difference between the same TIs and the average of the TIs for each type of cancer with embedded star graph
$\mathbf{F_{Score}}$	F-Score
$\mathbf{G_{mean}}$	G-mean
GP	Genetic Programming
Hb	Hydrophilicity
Hd	Hydrophobicity
IDME	Intelligent Decision Making Ensemble
IDMS-HBC	Intelligent Decision Making Ensemble for Diagnostic of Human Breast Cancer
KNN	K-Nearest Neighbor
$\mathbf{KNN_{AAC}/KNN_{SAAC}/KNN_{PseAAC\text{-}S}/KNN_{PseAAC\text{-}P}}$	AAC / SAAC / PseAAC-S / PseAAC-P feature space based KNN models
MCC	Mathew Correlation Coefficient
ML	Machine Learning
MTD	Mega-Trend Diffusion
MTD-SVM /KNN/…	SVM/… model using balanced data with MTD
PseAAC-P	Pseudo Amino Acid Composition-Parallel
PseAAC-S	Pseudo Amino Acid Composition-Series
PSO	Particle Swarm Optimization
pTIe	Cancer probability TIs with embedded star graph
QPDR	Quantitative Proteome-Disease Relationship
RF	Random Forest
RIA	Relative Improvement
ROC	Receiver Operating Characteristics
SAAC	Split Amino Acid Composition
Sn	Sensitivity
Sp	Specificity
SMOTE	Synthetic Minority Over-Sampling Technique

SVM	Support Vector Machines
SVM_{AAC}/SVM_{SAAC}/$SVM_{PseAAC-S}$/ $SVM_{PseAAC-P}$	AAC/SAAC/PseAAC-S/PseAAC-P feature space based SVM models
TIs	Topological Indices
TRR/FPR	True/False Positive Rate

Chapter 1: Introduction

The development of computational intelligent based decision making systems for the solution of a particular problem is an active area of research. The advancement in computational resources helps researchers to extract useful information from the vast amount of different kind of data. The development of intelligent decision making prediction systems that utilize the extracted information for the solution of a particular problem is critical. These systems have a broad range of applications and are being used for prediction in various fields of real life. The applications of prediction problems include: document prediction of spam e-mail messages, biometric identification (using fingerprints, iris patterns facial features, etc.), and cancer disease prediction using data of particular domain. The objective is to predict the existence of a particular discrete class (e.g., "cancer" versus \non-cancer" or spam vs. non-spam) based on the features of given instances of a certain dataset.

For healthcare systems, the development of an accurate and efficient decision support system has become an essential requirement. Such decision support system can provide adequate information for cancer diagnosis as well as drug discovery. In this thesis, the emphasis is on developing improved intelligent decision making ensemble classification systems (IDME) for accurate and efficient prediction of breast cancer. This chapter includes, background and motivation, ensemble system for prediction, objective and scope of work, contributions, and organization of the thesis.

1.1 Background and Motivation

Cancer is one of the rapidly growing diseases in the world [1]. Nearly, fourteen million people are diagnosed per year with cancer. It is estimated that this figure might increase up to 19 million in 2025. Additionally, it is assessed that out of 24 million cancer patients, half of them would be prevented in 2035 [2]. There are many types of cancers associated with human organs such as breast, colorectum, lung, and prostate. Commonly diagnosed cancers were related to lung (1.8 million, 13.0% of the total), breast (1.7 million, 11.9%), and colorectum (1.4 million, 9.7%) [3]. In case of breast cancer, about 1,300,000 new cases and 450,000 deaths are reported each

year worldwide. Breast cancer is a leading cause of death among women [4]. Approximately 1.5 million cases of women breast cancer are registered per year, worldwide. About 88% of women diagnosed with breast cancer would survive at least 10 years. In US, due to early prediction and treatment, more than 2.9 million women diagnosed with breast cancer were alive in 2012 [5]. In Pakistan, approximately 36750 new women breast cancer cases are estimated in 2015 and about 17552[1]women would die due to breast cancer. Like other cancers, breast cancer can also be successfully treated if predicted in the early stages. In order to increase the survival rate and to reduce cost, the early prediction of breast cancer with reliable and accurate decision support system is essential. Such an improved decision support system would be helpful to avoid unnecessary toxicity to the patient [6].

The rapid development in proteomics and genomics has resulted in the identification and generation of a large number of gene and protein sequences. This large amount of protein sequence data is being stored, and it could be utilized for the benefit of society such as for the early diagnosis, prediction of diseases, and drug discovery purposes. The use of protein sequencing is ever increasing and consequently, the number of protein databases is exponentially growing with time. Besides, the rapid increase in the protein databases is challenging task for extracting useful information for the prediction of cancer disease.

Molecular signatures of cancer are identified using fundamental knowledge of various disciplines of system biology, cell biology, structural biology, genomics, and proteomics. Using proteomics technology efforts are under way to develop molecular signatures/predictors of clinical outcomes and drug response [7, 8]. For diagnosis, treatment, and drug discovery of cancer different types of molecular predictors are needed. Macromolecules of proteins are composed of sequence of amino acids [9]. Protein molecular signatures are used for the prediction of ovarian cancer [10], lung cancer [11], colon cancer, and breast cancer [12]. In this research, molecular descriptors using the numerical values of physiochemical properties of protein amino acid sequences are exploited for the prediction of cancer. These properties have also

[1]*This figure is estimated using the population forecasts that were extracted from the United Nations, World Population prospects, 2012 revision (GLOBOCAN 2012).*

been used in the study of protein foldings, structures, protein-protein interaction, and sequence-order effects [13].

The prediction of breast cancer is an intricate process and specific indicators may produce negative results. Besides, conventional identification or diagnosis techniques such as mammography for breast cancer are time consuming, high-priced, and possess limited performance. In order to avoid misleading results, unbiased, effective, and efficient intelligent computational based decision support systems are always required by practitioners, and clinician so that patient could accurately be predicted for cancer. These systems could be helpful for increase the survival rate, reduce cost, and drug discovery [6]. For accurate prediction of cancer, the extraction of useful features from given dataset and then effective exploitation of their discriminative power is still a challenging task. The current research work has a strong motivation to help the humanity for their better life, and is targeted to a noble cause, i.e., the early prediction of breast cancer.

1.2 Ensemble System for Prediction

Breast cancer is a genomically complex and heterogeneous disease. Human body cells normally grow in a regular and controlled pattern. However, during cancer development, body cells grow without control. The cause of uncontrolled growth of abnormal cells is the mutation in genes. Cancer genomes are very unstable and usually show extensive genomic alterations. These changes vary from intragenic mutations to gross gains and/or loss of chromosomal material [14, 15]. Genetic mutations, deletions, and allelic loss of tumor suppressor genes may produce an aberrant Ribonucleic Acid (RNA) transcript that affects the normal function of the translated protein. Examples of commonly mutated tumor suppressor genes in human cancer are BRCA1 (breast cancer susceptibility gene 1), BRCA2, and P53 [16-18]. Usually most of the mutations that affect these genes result in protein sequences with truncations, small insertions or deletions. Such changes in protein sequences can be useful for cancer prediction and drug discovery. In this research, it is intended to exploit this discriminant change effectively in developing efficient and accurate prediction models.

The advancement in computational theories and tools allows scientists to extract useful information from rapidly increasing protein sequence datasets. Such

extracted information is utilizing by scientists to develop new prediction systems. These predictive systems may be developed using a single classification approach as well as a combination of multiple classification approaches. Systems employing a single approach are limited in terms of providing further/enhanced performance across the variety of datasets. Referring to classification problems, Wolpert's theorem has stated that no single optimal classifier exists for all pattern recognition problems because each classifier has its own domain of competence [3]. Consequently, there has been rising research interest in the approaches that combine diverse models to attain improved performance. There exist several combining approaches such as ensemble system, multiple classifier systems, committee of classifiers, or mixture of experts.

Early research has shown that ensemble system is usually more accurate than individual base predictor. The ensemble system effectively exploits the strengths of base-predictors, through the combination of their decisions, to improve performance. Thus, for a given prediction problem, it is expect that ensemble system well exploits the strengths of the base-predictors at our disposal to generate the high quality multiple recognition system overcoming the performance of base-classifiers. Ensemble systems have shown significantly better performance over other classification approaches and have provided powerful solutions to real life problems.

In ensemble learning, the basic presentiment is to construct a system by combining different hypothesis or predictive models. Technically, ensemble system refers to a collective decision making system that learns a target function by training a number of base classifiers/predictors (collection of decision makers) and applies a strategy to combine their decisions. Ensemble system has two main steps of ensemble generation and integration. In ensemble generation phase, accurate and diverse individual predictors are generated. A predictor is assumed an accurate if it has an error rate greater than random guessing on new instances. Two given predictors are supposed diverse if they commit distinct errors on new examples. On the other hand, ensemble integration involves the combination of decisions/predictions of individual predictors. In this manner, the learning of the base predictors enhances performance.

Ensemble system is a combination of computational models, which deal with all computational stages i.e., from data pre-processing to the final decision. Various

ensemble approaches have been proposed, but it is not straightforward to find an appropriate ensemble system for a particular problem. Fig. 1.1 shows different stages for the design of ensemble system for the prediction of cancer disease. In this figure, $C_1, C_2, ..., C_m$ represent base predictors, $h_1, h_2, ..., h_m$ correspond to their predicted values i.e., $h_i = C_i(\mathbf{x})$. The optimal/near-optimal hypothesis (h^*) is obtained using suitable ensemble strategy $F(h_1, h_2, ..., h_m)$. The unknown example $\mathbf{x}$ is predicted as cancer or non-cancer.

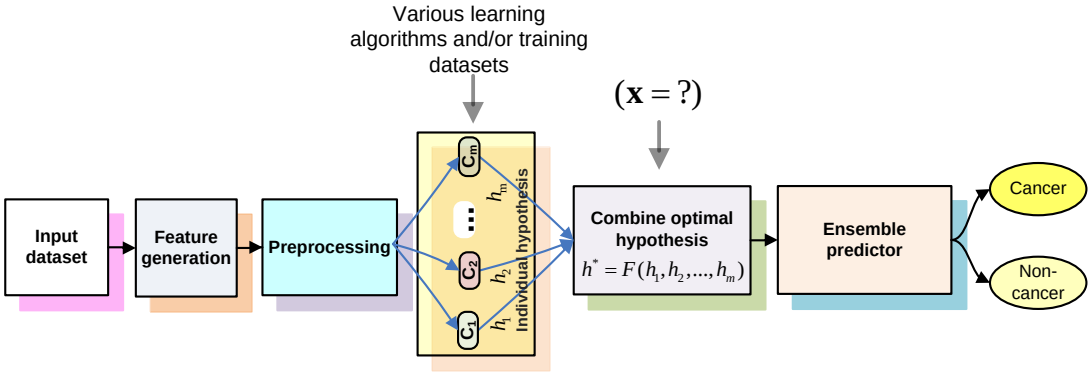


Figure 1.1 Various stages of a typical ensemble system for the prediction of cancer.

In the design of a decision making ensemble system, the following three fundamental issues must be considered to achieve better performance:

(a) Introduce adequate diversity

(b) Choose suitable base-predictors/classifiers

(c) Select proper ensemble strategy

Ensemble systems are either homogeneous or heterogeneous. Homogeneous ensemble system is constructed using single learning algorithm [24-26]. This ensemble system achieves diversity through some form of variability on the training dataset or injecting randomness into the parameters of the learning algorithm. However, Bagging [27] and Boosting [28] are data resampling based ensemble approaches. In contrast, the heterogeneous ensemble system is developed using diverse types of base predictors. Each slected base algorithm must displays a different inductive bias and learning hypotheses. The ensemble of predictors is generated using the training dataset. Their predictions/decisions are then combined through majority voting [19, 20], or more advanced methods [21, 22]. This type of ensemble system

attains diversity via different learning algorithms. Stacking is an example of the heterogeneous ensemble system [23].

1.3 Objective and Scope of Work

The primary objective of this research is the development of improved intelligent decision making ensemble systems for breast cancer prediction using protein amino acid molecules. The performance of such systems depends on various factors such as the nature of the problem, the selection of base learning algorithms, and the combining strategy. The performance depends on the nature of dataset as well, that may be balanced or imbalanced. These factors have certain effects on how accurately and efficiently the ensemble performs. This thesis seeks to develop various homogeneous and heterogeneous ensemble systems by employing different combining strategies. Further, the effect of imbalanced data on the performance of prediction systems is unveiled. The proposed systems are validated using cancer, non-cancer, and breast cancer datasets.

The thesis attempts to answer the following critical questions:

1. How can macromolecule of protein amino acid sequences be properly represented in feature spaces?
2. To explore optimal feature space(s) of protein sequences for cancer prediction?
3. How can the imbalanced data be handled for better prediction?
4. What types of combining strategy trainable or non-trainable can be adopted for improved performance of the ensemble system?
5. What are likely applications of this research?

1.4 Contributions of the Thesis

The worth of a research work is judged by the impact it has on society. The main contribution of the thesis is the development of novel high-performance IDME systems for the prediction of human breast cancer using protein amino acid sequences. The proposed IDME systems have shown improved performance over the existing approaches. Main contributions of this thesis are as follows:

1. Developing individual prediction systems (Chapter 3)

2. Applying data balancing and cost-sensitive learning techniques to cancer domain and unveiled the impact of imbalanced data on the performance (Chapters 3 and 4)
3. Developing a web based cancer predictor system (Chapter 4)
4. Developing homogeneous ensemble systems (Chapters 4 and 6)
5. Developing heterogeneous ensemble systems (Chapters 5 and 7)
6. Exploration of substantial variation in amino acid molecules of the cancerous proteins (Chapters 6 and 7)
7. Applying new molecular descriptors based feature spaces of protein amino acid sequences and exploration of optimal feature space(s) (Chapters 3-7)
8. Exploitation of physicochemical properties of amino acid molecules in different feature spaces (Chapters 3-7)

These contributions will be discussed in Chapters 3-7. The contribution presented in different chapters has been published in peer reviewed reputed journals. The list of these publications is given in the start of this thesis. The proposed systems can be used as a clinical decision support system for breast cancer prediction.

1.5 Thesis Organization

Fig. 1.2 highlights organization of this thesis. Chapter 2 begins with literature survey related to the development of ensemble systems. This chapter highlights various approaches in the literature developed for the prediction of breast cancer. In addition, different molecular descriptors based feature extraction strategies, adopted for feature generation from protein amino acid sequences, are discussed in detail. Description of various conventional approaches is presented. This chapter ends with the explanation of a number of performance measures utilized to evaluate the performance of the proposed systems.

Rest of the thesis describes various homogeneous and heterogeneous IDME systems. Chapters 3 and 4 highlight the basic framework for the development of improved systems and accentuate the exploration studies to handle imbalanced data. Chapter 3 presents the development of individual systems in the presence of imbalanced data. In this work, the issue of imbalanced data is explained. Development of various models, using different feature spaces, is discussed. Performance comparison of different models is also provided.

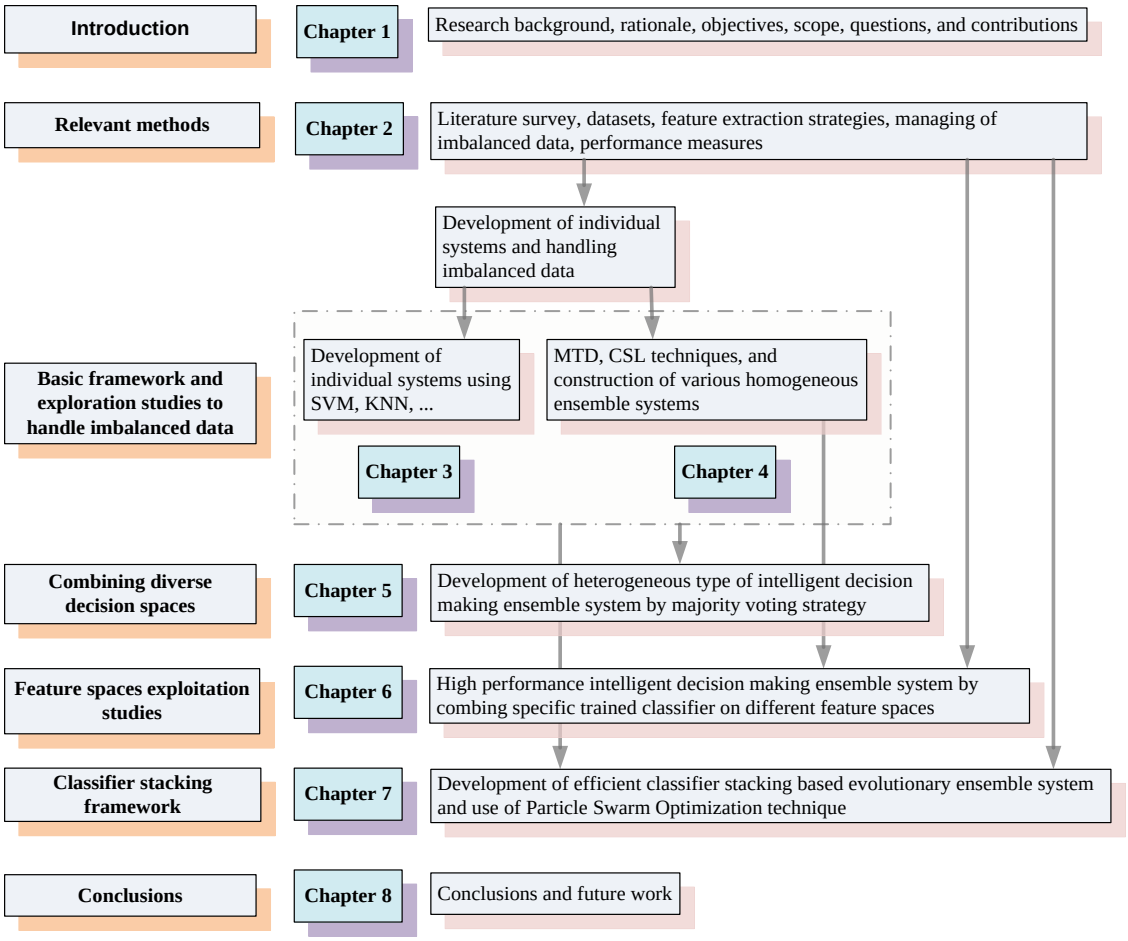


Figure 1.2 Thesis organization.

Chapter 4 elaborates the development of different homogeneous ensemble systems of Random Ensemble system and Cost-Sensitive Learning (CSL) based ensembles under balanced and imbalanced feature spaces. In this chapter, two novel approaches are discussed. These approaches are developed by combining Mega-Trend-Diffusion (MTD) with Random Forest (RF), and CSL technique with GentleBoost, AdaBoostM1, and Bagging. The performance of RF, GentleBoost, AdaBoostM1, and Bagging ensemble system is explored using both balanced and imbalanced feature spaces. It is found that MTD based ensemble system is superior than CSL based technique to handle imbalanced dataset of protein sequences. A web based cancer prediction system is also developed.

Chapter 5 contributes to the heterogeneous type of ensemble system. The improved heterogeneous Intelligent Decision Making Ensemble System for Human Breast Cancer (IDMS-HBC) is developed using several locally accurate and diverse predictors. This system showed better performance for breast cancer, particularly, in

case of PseAAC-S feature space. Overall, the proposed IDMS-HBC system outperformed over the individual, the conventional ensembles, and the previous approaches.

Chapter 6 accentuates on exploitation studies of feature spaces through the variation of amino acid molecules in cancerous protein sequences. It also devoted to elaborate the development of high performance Intelligent Decision Making Ensemble using PhysicoChemical properties of amino acids (IDM-PhyChm-Ens) systems. These ensemble systems are developed using diverse learning algorithms of RF, Support Vector Machines (SVM), and K-Nearest Neighbor (KNN)trained on different feature spaces.

The subsequent Chapter 7 presents the classifier stacking framework of Can-Evo-Ens system for efficient prediction of breast cancer. In this chapter, the Genetic Programming (GP) evolution process is discussed. This study revealed that PseAAC-S feature space has yielded excellent discrimination power over other feature extraction strategies. A comparative analysis of the evolutionary ensemble system with conventional ensemble approaches of AdaBoostM1, Bagging, and GentleBoost is also presented. Finally, Chapter 8 concludes this research work and sets the future directions.

Chapter 2: Related Work and Techniques

In this chapter, the conventional ensemble and previous approaches for the prediction of breast cancer are presented along with the detailed description of datasets of protein sequences used to assess the performance. Further, this chapter discusses various feature generation strategies, which are employed to develop intelligent prediction systems. The chapter concludes with the presentation of various measures, which are utilized to assess the performance of the intelligent systems.

2.1 Conventional Ensemble Approaches

Nowadays, ensemble systems have garnered an intense research interest in the machine learning community. One of the early approaches is the Jury Theorem that combined the results by applying misclassification probability for independent voters [29]. The idea of integration of approaches was first introduced by Chow [30] who set up rules for optimality of the combined decision of individual predictors with properly assigned weights. The first work for a classifier selection concept was introduced by Dasarathy and Sheela [31]. They proposed a strategy that jointed linear classifier and KNN classifier to distinguish the misclassification area in the feature space that considerably decreased the exploitation cost of the predictor system. However, a similar concept was independently developed by Rastrigin and Erenstein [32]. They separated a feature space into different regions and then individual predictor was trained on each region to attain the best classification accuracy. However, the hot wave of research on the development of ensemble system started since the 1990s. Hansen and Salamon [33] proposed a neural network ensemble, and Xu, et al [34] applied majority voting scheme for handwriting recognition. Turner and Ghosh [35] provided the analysis of decision boundaries in linearly combined neural predictors. Ho et al [36] presented work on decision combination in multiple predictor systems. They showed that a combined decision function must include useful information of base predictor.

Conclusively, the landmark success in the design of ensemble classifier is dedicated to introducing *Bagging* by Breiman [27] and *Boosting* by Schapire and

Freund [28, 37]. They were able to generate strong predictors from the weak one on the basis of theory of Probably Approximately Correct [38, 39]. Freund & Schapire proposed AdaBoost and its variant AdaBoostM1 and AdaBoostM2. Wolpert [23] worked on the concept of stacking (stacked generalization), and Ting and Witten [40] resolved two difficulties regarding the meta-classifier and its input features. Ordonez et al. [41] explored the optimizing ensemble configurations using Genetic Algorithms. Chen et al. [42] proposed Ant Colony Optimization based stacking approach by implementing the ideas of local information to accelerate the convergence of optimal solution.

2.2 Breast Cancer Prediction – Related Work

Several conventional, statistical, and computational intelligent techniques are used for the detection and prediction of breast cancer. Mammography is one of traditional methods to use for the detection of breast cancer; however, it has considerable variation in interpreting the results. Fine needle aspiration cytology is another conventional technique for the diagnosis of breast cancer with the limited identification rate of 90%. For this reason, in recent times, many statistical, computational, and intelligent approaches are being considered for the prediction of breast cancer. In a recent study, benign and malignant breast lesions are differentiated using 3D Magnetic Resonance Imaging (MRI) morphological features based model [43]. Accurate methods are attempted using graph theory for Deoxyribonucleic Acid (DNA) sequence analysis, prediction of protein properties, and prediction of breast and colon cancers. The protein quantitative structure-activity relationship [44] and Quantitative Proteome-Disease Relationship (QPDR) [45] are used for the prediction of protein properties and cancers.

Multi-target QPDR cancer prediction models are developed using macromolecular descriptors of Topological Indices (TIs), which were obtained from the graph theory [12]. The proposed QPDR models are based on multiple linear regression technique. Due to nature of multiple linear regression models, it is difficult to model accurately the nonlinearity in the features to target labels. Therefore, the maximum performance of QPDR models is limited to 91.80% for cancer prediction.

On the other hand, in the contemporary literature, several feature extraction strategies and computational intelligence approaches have been used for the

development of cancer decision support systems [25, 46-50]. Role of feature extraction is very crucial in the development of a reliable and effective method for the construction of diagnosis systems. Extracted feature vectors are used by such systems to establish the classification model. A good feature set is supposed to be highly correlated within a class and uncorrelated with other classes. The main objective of these approaches was to improve the performance with hybrid or ensemble systems. For breast cancer prediction, gene expression features were selected using Genetic Algorithm and SVM nonparallel plane proximal classifier was developed to obtain the accuracy of 88.21% [51]. The recurrence modeling based feature extraction strategy has reported the prediction accuracies up to 93.60% and 94.70 % using Decision Tree (DT) and Artificial Neural Networks (ANN), respectively [52]. In other studies, accuracy values of 94.10% using Association Rules and Neural Network [53], 94.74% using C4.5 DT method [54], and 95.06% using Neuron-Fuzzy methods [55] were obtained. In recent study, for the assessment of estrogen receptor status in breast cancer tissue, several predictors of SVM, Radial Basis Function Network, KNN search, Naïve Bayes (NB), Functional Trees, and K-means clustering algorithm were applied [56]. In another study, multiple sonographic and textural features based methods were developed for the classification of benign and malignant breast tumors [57].

The RotBoost ensemble system applied using microarray genes and reported 94.39% accuracy [58]. This ensemble system is constructed by integrating the ideas of Rotation Forest and AdaBoost. In another study, RF ensemble system was used with sequence-based features for the prediction of DNA-binding residues in protein sequences [59]. The Bayesian Network technique was applied to construct a gene regulatory network from microarray data and achieved an Area Under Curve of the Receiver Operating Characteristics (AUC-ROC) of 79.20% [60]. In another work, simulated and real breast cancer datasets were employed to evaluate the performance of ANN models [61]. Several breast cancer studies were performed using clinical features extracted from the image of a fine needle aspirate (FNA) of a breast mass. For instance, these clinical features were used for non-evolutionary method Optimized-LVQ (Learning Vector Quantization), Big-LVQ, and Artificial Immune Recognition System (AIRS) [62]. Evolutionary Algorithms (EAs) with the combination of SVM using clinical features from FNA of breast mass provided

accuracy up to 97.07% [63]. The EA-based feature selection approaches achieved improved performances [64]. Protein features for malignant and benign cancers were evaluated using different screening methods, for example, DT models, generalized rule induction, and clustering methods for identification of similarity patterns in benign and malignant breast cancer tissues [65]. In another study, Bagging and Boosting ensemble approaches were used for the decision making system [66]. These ensemble systems used Classification and Regression Trees for feature selection and achieved prediction accuracies up to 97.85 % and 95.56%, respectively.

Usually, in cancerous data, number of cancer and non-cancer patients is inherently imbalanced i.e., the particular diagnosis class is not easily achievable. Several prediction models are developed to handle imbalanced problem using processing the input data by over/under-sampling or cost sensitive learning (CSL) technique. Prediction models were developed using DT for breast cancer survivability and showed the 86.52% survival rate of patients [67]. This work dealt to handle imbalanced problem using the under-sampling approach and hence improved the prediction performance. In a recent study, Linear Regression (LR) and DT has been used to develop models for the five-year prognosis [68]. These models were constructed by combining Bagging, Boosting, the synthetic minority over-sampling technique (SMOTE), under-sampling, and cost sensitive classfier (CSC) techniques. This study showed that DT and LR based models with SMOTE, under-sampling, and CSC techniques produce better results than imbalanced datasets. These models achieved accuracy up to 91.30%. In an another study, prediction models were developed using under-sampling approach with DT to deal with imbalanced problem for breast cancer survivability and reported 86.52% survival rate of patients [67]. Models were constructed utilizing gene expression profiles for the prediction of breast cancer by employing LR, SVM, AdaBoost, LogitBoost and RF [69]. They achieved maximum AUC of 88.6% and 89.9% for SVM and RF models, respectively. In an another study, surveillance and epidemiology results were used the for prediction of breast cancer [70]. This study employed three classification models of DT, LR, and ANN and reported highest AUC of 84.9% and 76.9% for LR and DT, respectively. In a recent study, prediction models were developed from highly imbalanced data using SVM, Bagging, Boosting and RF [71]. Results demonstrated (AUC) that, RF model (91.2%) outperformed SVM (90.6%), Bagging (90.5%), and Boosting (88.9%).

This discussion highlights that though several prediction systems are proposed, it is not an easy task to obtain improved ensemble system for a specific dataset. Generally, researchers have either developed novel approaches or suggested modifications to the existing methodologies. The next section describes the datasets of protein amino acid sequences that are utilized during the course of this research.

2.3 Datasets of Protein Primary Sequences

Healthy and disease-samples of cancer are predicted using datasets of protein amino acid sequences. There are twenty naturally-occurring amino acids. These different types of amino acids have shown chemical versatility and performed a wide variety of functions in the body. They work as the necessary constituents of body tissues, enzymes, and immune cells. The most important function of amino acids is to act as the building blocks of protein molecules. The sequence of amino acids in a protein is obtained using genetic codes in the DNA. The native amino acids in a protein sequence are usually illustrated by a set of single letter codes of English letters. A protein of length L is formally represented as an ordered sequence $p = (a_1, a_2, \ldots, a_L)$ with elements a_i, from the finite set $= \{\text{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}\}$, where A stand for Alanine, C stand for Cysteine, E stand for Glutamic acid, etc. Examples of non-cancer (Seq. 1), cancer (Seq. 2), and breast cancer (Seq. 3) protein amino acid sequences in FASTA format are given below:

Seq.1:
1A0S::PSGFEFHGYARSGVIMNDSGASTKSGAYITPAGETGGAIGRLGNQAD...

Seq.2:
ADAMTS15::MLLLGILTLAFAGRTAGGSEPEREVVVPIRLDPDINGRRYYW…

Seq.3:
ABCA3::MAVLRQLALLLWKNYTLQKRKVLVTVLELFLPLLFSGILIWLRLKI...

where A, C, E, … , be the standard single letter codes of different native amino acids in protein sequence.

The datasets of proteins primary sequences are represented in composite protein sequences. A composite protein sequence consists of 20 unique amino acids. Owing to differences in side chains, all amino acids have different chemical

properties, but possess a common basic chemical structure. A composite protein sequence could be represented by a chain of amino acids. Different proteins have different amino acid strings, in terms of the ordering and the total length of the sequence.

Performance of the proposed systems is reported using datasets of protein amino acid sequences of Cancer/Non-Cancer (C/NC), Breast/Non-Breast Cancer (B/NBC), and Colon/Non-Colon Cancer (CC/NCC). These datasets are taken from Sjoblom and Dobson group [72, 73]. These groups have extracted cancer-related proteins after the experimental analysis of 13,023 genes in 11 breast and 11 colorectal cancers. The first dataset has 865 non-cancer protein sequences. The other datasets have 122 breast-cancer and 69 colon-cancer related protein sequences. In this way, 191 cancer protein sequences are obtained. Table 2.1 shows the nature of C/NC, B/NBC, and CC/NCC datasets. The imbalance ratio is obtained by dividing the number of negative examples (not disease) by the number of positive examples (disease). A dataset is called balance if the imbalance ratio is equal to one. The majority to minority class ratios have values of 4.53, 7.66, and 14.30 for datasets of C/NC, B/NBC, and CC/NCC, respectively. Dataset of the minority class of CC/NCC is relatively highly imbalanced.

Table 2.1 The nature of imbalanced datasets.

Dataset	Positive examples	Negative examples	Imbalance ratio
C/NC	191 (19.09%)	865 (81.91%)	4.53
B/NBC	122 (11.55%)	934 (88.45%)	7.66
CC/NCC	69 (6.53%)	987 (93.45%)	14.30

2.4 Feature Generation Strategies

Role of feature extraction is very crucial in the development of a reliable and effective decision making system. Extracted feature vectors are used by systems to establish the prediction model. A good feature set is supposed to be highly correlated within a class and uncorrelated with other classes. Proper input representation of protein primary sequences make easier for a classifier to recognize underlying regularities in sequences. The raw information given by the protein sequence is customarily restructured for prediction such that each representation of protein sequences is suitable for a feature space. For cancer prediction, molecular descriptors derived from physicochemical properties of amino acids are used to transform primary protein

sequences into feature spaces of amino acid composition (AAC), split amino acid composition (SAAC), pseudo amino acid composition-series (PseAAC-S), and pseudo amino acid composition-parallel (PseAAC-P). Table 2.2 shows the feature spaces along with their dimensions. Following paragraphs explain these feature generation strategies.

Table 2.2 Dimension of different feature spaces.

Sr. No.	Feature Space	Dimension
1	Amino acid composition	20
2	Split amino acid composition	60
3	Pseudo amino acid composition-Series	$60^{\#}$
4	Pseudo amino acid composition-Parallel	$40^{\#\#}$

$^{\#}20 + \lambda$ and $^{\#\#}20+i*\lambda$ where λ is tiers level and 'i' is number of amino acid attributes, here $\lambda=20$ and $i=2$ are used.

2.4.1 Amino Acid Composition

In AAC feature space (20-dimensions), each protein is represented by a vector of the relative frequencies of 20 native amino acids in its sequence as:

$$f_i = \frac{n_i}{L} \qquad (i = 1, 2, ..., 20) \tag{2.1}$$

where f_i represents the occurrence frequency of the ith native amino acid in the protein, n_i is the number of ith native amino acid in sequence. Then, the AAC feature vector is expressed as:

$$\mathbf{x}_{AAC} = \left[f_1, f_2, ..., f_{20} \right]^T \tag{2.2}$$

2.4.2 Split Amino Acid Composition

In the SAAC feature space generation, the given protein amino acid sequence is split into three dissimilar sections, named the N-terminal, the Internal segments and the C-terminal [74]. Fig. 2.1 shows the partition of the protein sequence into N-terminal, Internal segments and C-terminal for SAAC feature space. The amino acid compositions of three parts are computed individually using Equations 2.1 and 2.2. The length of each part is set equal to 20. In this way, the dimension of each feature vector in the SAAC model becomes 60. The SAAC feature vector is given as:

$$\mathbf{x}_{SAAC} = \left[nt_1, \ldots nt_{20}, is_1, \ldots, is_{20}, ct_1, \ldots, ct_{20} \right]^T \tag{2.3}$$

where $\mathbf{N}=[nt_1,\ldots,nt_{20}]^T$, $\mathbf{I}=[is_1,\ldots,is_{20}]^T$, and $\mathbf{C}=[ct_1,\ldots,ct_{20}]^T$ are the N-terminal, the Internal segments and the C-terminal feature vectors of a protein.

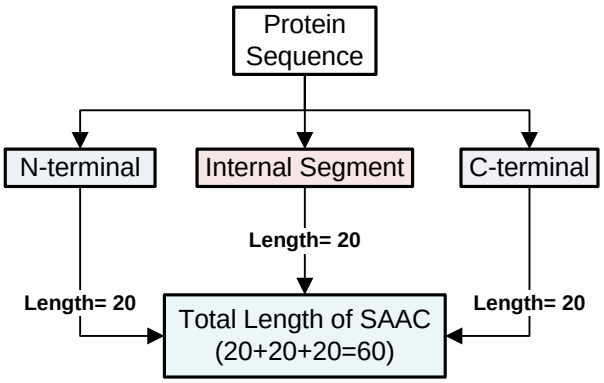


Figure 2.1 Construction of SAAC features.

2.4.3 Puesdo Amino Acid Composition

The PseAAC based feature spaces have capability to carry core and essential information concealed in complicated protein sequences without losing its sequence-order. Two types of series and parallel PseAAC correlations based features are generated. In the series correlation (PseAAC-S), a protein feature vector is generated using $20+i\times\lambda$ discrete components, where '*i*' is the selected number of amino acid properties. The protein vector in the series correlation of dimensions 60, with *λ=20* and *i=2* for hydrophobic and hydrophilic values, is expressed as:

$$\mathbf{x}_{PseAAC-S}=\left[p_1\cdots p_{20}\,p_{21}\cdots p_{20+\lambda}\,p_{20+\lambda+1}\cdots p_{20+2\lambda}\right]^T \quad (\lambda<L) \tag{2.4}$$

with

$$p_u=\begin{cases}\dfrac{g_u}{\sum_{i=1}^{20}g_i+w\sum_{j=1}^{2\lambda}\tau_j} & \text{for} \quad 1<\mu<20\\[2ex] \dfrac{w\tau_u}{\sum_{i=1}^{20}g_i+w\sum_{j=1}^{2\lambda}\tau_j} & \text{for} \quad 21<\mu<20+2\lambda\end{cases} \tag{2.5}$$

where, $g_i\left(\mathrm{i}=1,2,\ldots,20\right)$ be the normalized occurrence frequencies of 20 native amino acids in the protein and τ_j the *j*th-tier sequence-correlation factor is calculated according to $\tau_{2\lambda-1}=\dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}Hd_{i,i+\lambda}$ and $\tau_{2\lambda}=\dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}Hb_{i,i+\lambda}$, L is the amino acid residues. Here, a weighting factor *w* is empirically set equal to 0.05.

The protein feature vector in parallel correlation (PseAAC-P) is given by $20+\lambda$ discrete components as:

$$\mathbf{x}_{PseAAC-P} = \left[p_1 \ldots p_{20} p_{20+1} \ldots p_{20+\lambda} \right]^T \tag{2.6}$$

with

$$p_u = \begin{cases} \dfrac{g_u}{\sum_{i=1}^{20} g_i + \omega \sum_{j=1}^{\lambda} \theta_j} & \text{for} \quad 1 \le \mu \le 20 \\ \dfrac{\omega \theta_{u-20}}{\sum_{i=1}^{20} g_i + \omega \sum_{j=1}^{\lambda} \theta_j} & \text{for} \quad 21 \le \mu \le 20+\lambda \end{cases} \tag{2.7}$$

where, θ_λ be the λ-tier correlation factor that reveals the sequence order correlation between all the λ for the most contiguous residues along a protein chain. It is determined using the following equation.

$$\theta_\lambda = \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} \Theta\left(R_i, R_{i+\lambda}\right) \tag{2.8}$$

Thus, for $\lambda=20$, a feature vector of 40 dimensions is generated. The value of $\Theta\left(R_i, R_j\right)$ is computed as follows:

$$\Theta\left(R_i, R_j\right) = \frac{1}{2}\left\{ \left[H_d\left(R_j\right) - H_d\left(R_i\right) \right]^2 + \left[H_b\left(R_j\right) - H_b\left(R_i\right) \right]^2 \right\} \tag{2.9}$$

where, Hd and Hb are hydrophobicity and hydrophilicity of *i*th amino acid R_i and *j*th amino acid R_j, respectively. The Hd and Hb values used are given in Table 2.3.

2.5 Conventional Computational Approaches

Various conventional individual and ensemble approaches based decision making systems are developed using different feature space of AAC, SAAC, PseAAC-S, and PseAAC-P for the prediction of cancer. The most common individual approaches are KNN, SVM, NB, DT, and Probabilistic Neural Network (PNN). Beside these computational approaches conventional ensemble approaches of RF, Bagging, Boosting, AdaboostM1, and Gentleboost are utilized. GP is also employed to develop evolutionary ensemble system. These learning algorithms are implemented in MATLAB R2013a environment. A brief description of these computational approaches with respect to implementation is given. The detail description is available in the literature of computational intelligent.

Table 2.3 Single letter codes and numerical values of Hd and Hb of twenty native amino acid molecules of protein.

Sr. No.	Amino acid	Single letter code	Hd [75]	Hb [76]
1.	Alanine	A	0.62	-0.5
2.	Cysteine	C	0.29	-1
3.	Aspartic acid	D	-0.9	3.0
4.	Glutamic acid	E	-0.74	3.0
5.	Phenylalanine	F	1.19	-2.5
6.	Glycine	G	0.48	0
7.	Histidine	H	-0.4	-0.5
8.	Isoleucine	I	1.38	-1.8
9.	Lysine	K	-1.5	3.0
10.	Leucine	L	1.06	-1.8
11.	Methionine	M	0.64	-1.3
12.	Asparagine	N	-0.78	0.2
13.	Proline	P	0.12	0
14.	Glutamine	Q	-0.85	0.2
15.	Arginine	R	-2.53	3.0
16.	Serine	S	-0.18	0.3
17.	Threonine	T	-0.05	-0.4
18.	Valine	V	1.08	-1.5
19.	Tryptophan	W	0.81	-3.4
20.	Tyrosine	Y	0.26	-2.3

2.5.1 Individual Learning Approaches

In this section, implementation details of KNN, SVM, NB, DT, and PNN approaches are explained.

K-Nearest Neighbor

KNN approach is mostly used in prediction studies. It is simple in implementation and provides effective performance. Here, Euclidean distance is considered as a metric to measure the proximity. If classification is based on labels of more than one nearest located examples, then it is known as KNN classifier. K indicates the number of neighbors to be located nearby the query protein.

For a query protein sequence **x**, how can one predict its class label? According to the nearest neighbor principle, we have to find generalized distance S between **x**

and $\mathbf{x}_i$:

$$S(\mathbf{x}, \mathbf{x}_i) = 1 - \frac{\mathbf{x}.\mathbf{x}_i}{\|\mathbf{x}\|\|\mathbf{x}_i\|} \quad \text{(i=1,2,3,…,N)} \tag{2.10}$$

where, $\mathbf{x}.\mathbf{x}_i$ is the dot product of vectors, $\mathbf{x}$ and $\mathbf{x}_i$; and $\|\mathbf{x}\|$ and $\|\mathbf{x}_i\|$ are respectively their magnitudes. The generalized minimum distance is calculated as:

$$S(\mathbf{x}, \mathbf{x}_k) = Min\{S(\mathbf{x}, \mathbf{x}_1), S(\mathbf{x}, \mathbf{x}_2), \cdots, S(\mathbf{x}, \mathbf{x}_N)\} \tag{2.11}$$

The query protein sequence $\mathbf{x}$ is assigned the category corresponding to the training protein $\mathbf{x}_k$. During the implementation of KNN, different values 1, 3, 5, … of K are tried.

Support Vector Machines

SVM model is well documented in the statistical learning theory [77]. It is a popular learning algorithm used to perform various prediction tasks in bioinformatics. Here, same basic notations of SVM theory to establish the equation of hyperplane are used. For a linearly separable data of N training pairs (x_i, y_i), the function of a decision surface is defined as:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i^T . \mathbf{x} + bias \tag{2.12}$$

where, the coefficient $\alpha_i > 0$ is the Langrange multiplier in an optimization problem. The pattern vector $\mathbf{x}_i$ corresponds to $\alpha_i > 0$ is called a support vector (SV). In order to find an optimal hyperplane surface for non-separable patterns, solution of the following optimization problem is sought:

$$\Psi(\mathbf{w}, \zeta) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\zeta_i \tag{2.13}$$

subject to the condition $y_i \left(\mathbf{w}^T \Psi(\mathbf{x}_i) + bias\right) \geq 1 - \zeta_i, \quad \zeta_i \geq 0$

where, *C* is the penalty parameter of the error term $\sum_{i=1}^{N}\zeta_i$. It represents the cost of constraint violation of those data point, which occurs on the wrong side of the decision boundary and $\Psi(\mathbf{x})$ is the nonlinear mapping. The weight vector $\mathbf{w}$ minimizes the cost function term $\mathbf{w}^T\mathbf{w}$. The nonlinear decision surface can now

be constructed by a new function as:

$$f(\mathbf{x}) = \sum_{i=1}^{N_S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + bias = \sum_{i=1}^{N_S} \alpha_i y_i \Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x}) + bias \qquad (2.14)$$

where, Ns represent the number of SVs. For SVM decision function, radial basis function with Gaussian kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2}{(2\sigma)^2} \right)$ is used, where parameter σ shows the width of Gaussian kernel.

During parameter optimization, the optimal values of *C* and σ are obtained with the grid search method, respectively. The number of support vectors is greater than 10% for the original imbalance dataset of proteins, consequently inaccurate performance due to overfitting of the data is obtained. During the SVM training phase, several experiments are performed with the varying size of reference balanced datasets, i.e., add diffuse/synthetic samples in original datasets. It is observed that the best performance with the balanced data is achieved. The best SVM model is selected, for the number of support vectors, about 10% of the reference dataset.

Naïve Bayes

NB is a statistical classification method. It is an efficient and effective inductive mechanism for computational learning. During training, NB uses methods based on the Bayes'Theorem (Eq.(2.15)) to estimate the probability of relating certain classes at certain examples given the values of the predictor variables.

$$P(c_{li} \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_{li}) P(c_{li})}{P(\mathbf{x})} \qquad (2.15)$$

NB predictor could classify an unseen example of a certain class by assuming its features are conditional independent [78]. Let C_l is the random variable representing the class of an example i.e., label and $\mathbf{x}$ is a protein vector of random variables representing the observed features values. Let c_l is a certain class label and x is denoting a certain observed feature value. From training dataset, the conditional probability of each $\mathbf{x}_i$ provided the class label C_l is learned. Classification is performed by employing Bayes rule to calculate the probability of C_l for $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n$, using the following Eq. as:

$$P(C_l = c_l \mid \mathbf{x}_1 = x_1, ..., \mathbf{x}_n = x_n) \quad (2.16)$$

Because protein features are assumed independent, the posterior probability of the class is given as follows:

$$P(C_l = c_l \mid \mathbf{x}_1 = x_1, ..., \mathbf{x}_n = x_n) = P(C_l = c_l) * \prod_{i=1}^{n} P(\mathbf{x}_i = x_i \mid C_l = c_l) \quad (2.17)$$

For two classed dataset, the predictor predicts the class which has the highest posterior probability:

$$\max_{c_l} \prod_{i=1}^{n} P(\mathbf{x}_i = x_i \mid C_l = c_l) \quad (2.18)$$

Decision Tree

DT is a simple and powerful technique for prediction. Several types of algorithms are used to construct DT models [79]. To construct a tree structure, DT recursively divides data examples into branches. Each tree node is either a decision node or leaf node. Decision nodes are split according to the given criterion which testing the values of certain functions of data features. Distinct outcome of the test generates branches of the decision node. Fig. 2.2 shows class label attached to each leaf node. When training examples at the *n*th node are of the same class label c_l , then this node

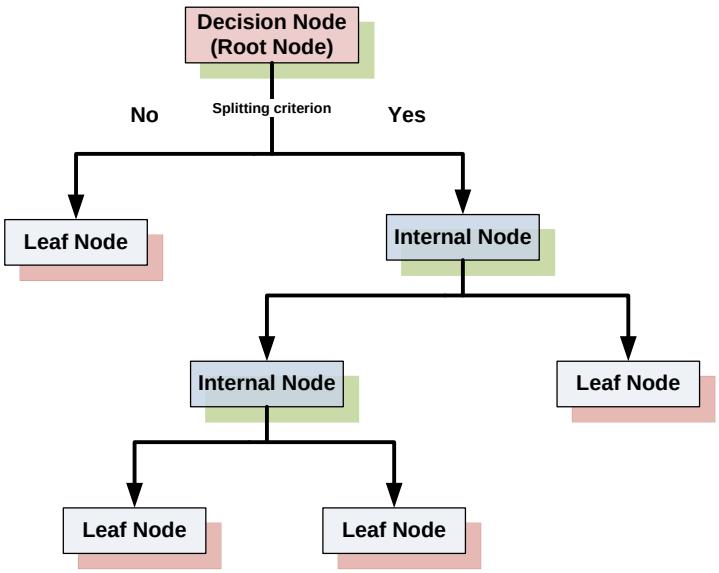


Figure 2.2 An example of binary decision tree.

develops into a leaf node with label c_l. Otherwise, pick the splitting of the most important features in separating the training examples into different classes. This

feature develops into a decision node. This procedure carry on until a particular stopping criterion is fulfilled.

Probabilistic Neural Network

PNN was proposed by Specht [80]. It is based on Bayes classification rule and is used to construct probability density function for each classification category using nonparametric estimation theory called Parzen windows. This learning algorithm is implemented with kernel discriminant analysis. PNN is an arrangement of various interconnected processing units or neurons structured in successive layers. It is constructed with four layers; the input layer, the pattern layer, the summation layer, and the output layer, as shown in Fig. 2.3.

No computation is performed in the input layer unit; it simply distributes the input vector $\mathbf{x}$ to all neurons in the pattern layer. The neurons of the pattern layer are separated into j groups, one for each class. The neuron, say $\mathbf{x}_{ji}$, *j*th pattern node of the *i*th class, calculates its probability density function as output using the following function:

$$\psi_{ij}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^{d}} \exp\left(\frac{-\|\mathbf{x}-\mathbf{x}_{ij}\|}{2\sigma^{2}}\right) \tag{2.19}$$

where, $\mathbf{x}$ is the input pattern vector, d is the dimension of protein feature space, and σ is smoothing parameter, which determines the width of activation function. Training data examples in *i*th class out of m classes is represented by N_i. The output of the pattern layer is presented to the summation layer. The summation layer of the network computes maximum likely-hood of $\mathbf{x}$ being classified into class c_l through a combination of previously computed densities.

$$p_{ri}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^{d}} \frac{1}{N_i} \sum_{j=1}^{N_i} \exp\left(\frac{-\|\mathbf{x}-\mathbf{x}_{ij}\|}{2\sigma^{2}}\right), \quad \text{for } i = 1,\ldots,m \tag{2.20}$$

Here, we set m=2 for our binary class problem. The output of the summation layer is feed to the output layer or decision layer of the network. The output layer constructs the final decision of the class for the data pattern by employing the optimal Bayes decision rule.

$$c_l(\mathbf{x}) = \text{Arg}\max\left\{p_{ri}(\mathbf{x})\right\} \qquad \text{for } i = 1,2 \tag{2.21}$$

where, $c_l(\mathbf{x})$ represents the predicted class for the input pattern $\mathbf{x}$.

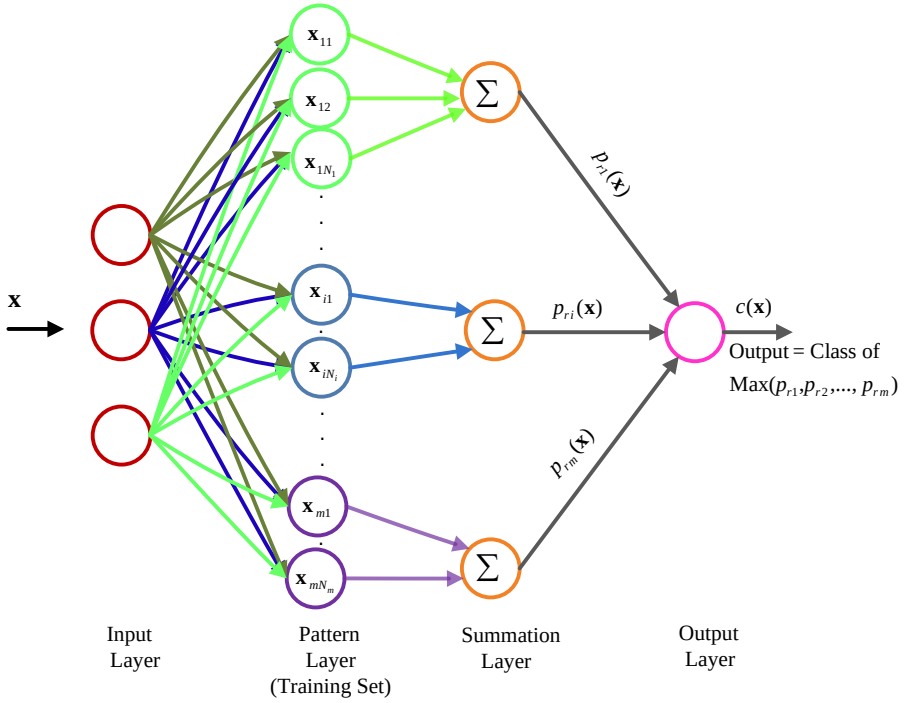


Figure 2.3 Working architecture of PNN approach.

2.5.2 Ensemble Based Learning Approaches

In this section a brief description of RF, Bagging, AdaBoostM1, GentleBoost, and GP is presented.

Random Forest Ensemble

RF algorithm uses an ensemble of classification trees [81]. It is being used to solve different types of problems in bioinformatics. It has performed well with small data size but complex data structures. It is a successful approach for combining unstable predictors [82], and random variable selection for tree building.

Working procedure of the RF algorithm is shown in Fig. 2.4. It uses two random mechanisms (i) random feature selection and (ii) bootstrap data sampling. RF utilizes random sampling and ensemble properties that make it possible to attain accurate prediction and better generalization. For n features, a number $mtry<n$ is specified such that at each node, *mtry* out of n features are randomly selected. The value of *mtry* is selected to be $\sqrt{n}$. During forest growing, the value of *mtry* remains constant. The best split on these *mtry* features is used to split the node. Gini measure is used as a splitting criterion that selects the split of each node with the lowest impurity. Gini measure of impurity is given as follows.

$$\text{Gini}(t,\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{j} a_i I(t_{x_i}) \tag{2.22}$$

where, *j* is the number of children at node *t*, *N* is the number of samples, $I(t_{x_i})$ is Gini impurity, which gives the class label distribution in the node. $I(t_{x_i})$ for the variable $\mathbf{x} = \{x_1, x_2, ..., x_j\}$ at node *t* is given by:

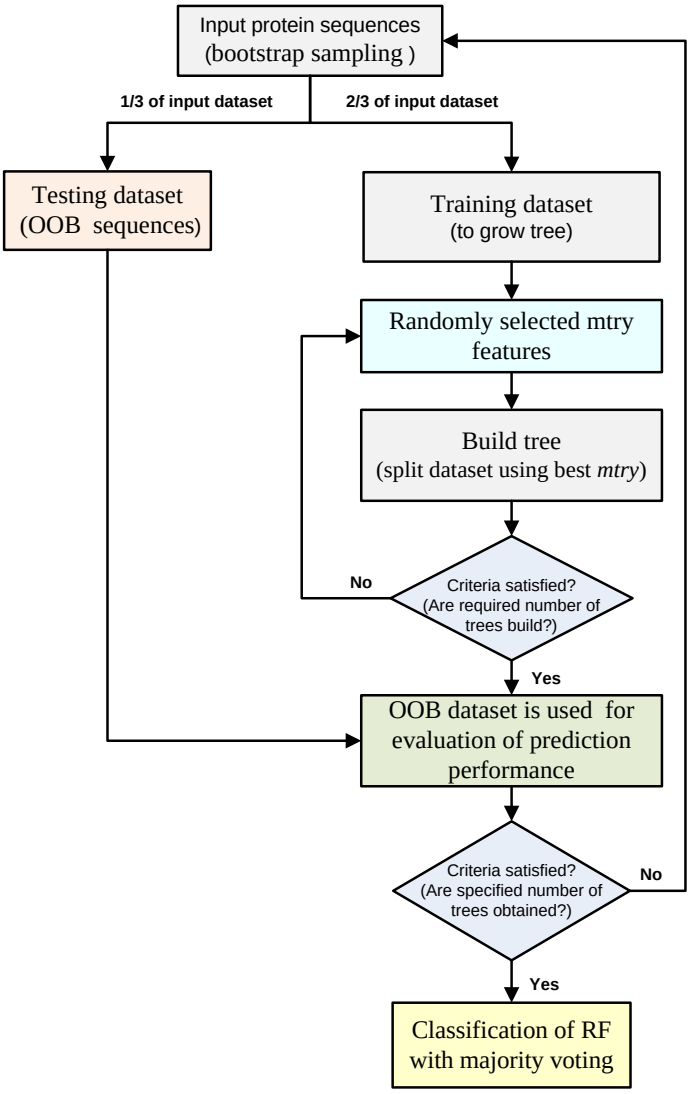


Figure 2.4 Block diagram of the RF ensemble system.

$$I(t_{x_i}) = 1 - \sum_{c_l=0}^{c_l} \left(\frac{n_{c_{li}}}{a_i} \right)^2 \tag{2.23}$$

where, $n_{c_{li}}$ is the number of samples with value x_i belonging to class c_l, a_i is the number of samples with value x_i at node t. Consider a dataset *D* of *N* protein

sequences for training: $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, where $\mathbf{x}_i$, $i = 1, 2, 3, ..., n$, is *i*th protein vector corresponding to y_i target label. The output predicted values, for a given $\mathbf{x} \in D$, are given by $\left\{\hat{y}_1(\mathbf{x}), \hat{y}_2(\mathbf{x}), \ldots, \hat{y}_T(\mathbf{x})\right\}$, where $\hat{y}_t(\mathbf{x})$ is the predicted value of protein sequence $(\mathbf{x})$ by the *t*th tree. Outputs of all predicted values are aggregated using majority vote for final decision.

Bagging, AdaBoost, and GentleBoost

Bagging (**b**ootstrap **agg**regating) is a data resampling based ensemble approach [27]. This approach utilizes the idea of bootstrap samples to develop diverse base predictors. The bootstrap sample of *m* examples is created by uniformly sampling with replacement from the training dataset. The bootstrap sample is contained approximately (1-e^{-1}) 63.2% different examples. The ensemble system generates T bootstrap samples $B_1, B_2, \ldots, B_T$ and corresponding T base predictors $C_1, C_2, \ldots, C_T$ are built. For final decision, predictions of *T* base predictors are then combined by the majority voting scheme. In case of base predictor is unstable, Bagging facilitates to decrease the error related to random variations in the training dataset. But in case of base predictor is stable, the sources of error of the ensemble is mainly due to bias in the base predictor. Bagging ensemble is implemented using discriminant analysis as base learning algorithm.

AdaBoost (**Ada**ptive**Boost)** is an improvement of the early version of boosting algorithm [83]. AdaBoost is a special case of sequential topology. It has adaptability of the succeeding distributions to the output of the previous weak predictors. It modifies the weights of the training examples following every assessment based on the misclassifications of base predictor and induce the learning algorithm to reduce the expected error. The correctly predicted examples have assigned lower weight. Therefore, for weighted examples, AdaBoost produces *T* training datasets $S_1, S_2, \ldots, S_T$ during *T* assessments and then the corresponding *T* base predictors $C_1, C_2, \ldots, C_T$. There are several variants on the idea of Boosting. AdaBoostM1 is designed specifically for binary prediction.

Friedman et al. modified the Real AdaBoost algorithm to develop GentleBoost algorithm. Generally, GentleBoost outperforms Real AdaBoost at stability. This

ensemble approach is chosen because it is numerically robust and performs better over other boosting variants in various domains of applications. More details about GentleBoost ensemble are available in [84]. In this algorithm, the optimization of cost function is performed by employing adaptive Newton steps, which keeps to decreasing weighted squared error at every step. AdaBoostM1 and GentleBoost ensemble approaches are implemented using DT as base predictors.

Genetic Programming

GP technique is based on the principles of natural selection and recombination under defined fitness criterion [85]. It is a powerful evolutionary approach, which solves problems automatically and searches for possible solutions in the defined problem space. This approach was used effectively in different applications of pattern recognition [86-89]. GP is employed to develop a new ensemble system of ameliorated performance for breast cancer prediction. During GP evolution process, predictions of the base-level predictors are combined using the fitness criterion of AUC-ROC. The GP evolution process is discussed in detail in chapter 7.

2.6 Performance Measures

Performance evaluation is a key phase during the development of a prediction system. In Machine Learning, various proficient performance measures are recognized for the performance evaluation of algorithms. However, in various problem domains the use of a specific performance measure is mostly based on the classification task and data distribution of the domain of interest. Performance measures are derived from a confusion matrix. This matrix is tabulated the actual labels versus the predicted labels for each class. For a binary class problem such as cancer (Positives) and non-cancer (Negatives) a confusion matrix is represented in Table 2.4. In this Table, TP (true positives) specifies the number of positive examples predicted as positives whereas FN (false negatives), also known as error of the second kind, signifies the number of positive examples predicted as negatives. In the same way, FP (false positives), also known as error of the first kind, indicates the number of negative examples predicted as positives and TN (true negatives) shows number of negative examples predicted as negatives. These measures are based on the correct and wrong prediction of the predictor. The popular performance measures, which are extensively employed by ML and pattern recognition community, are given below.

Table 2.4 Confusion matrix for binary problem.

		Predicted class	
		Cancer (Positives)	Non-Cancer (Negatives)
Actual class	Cancer (Positives)	TP	FN
	Non-Cancer (Negatives)	FP	TN

2.6.1 Accuracy

Accuracy (or error rate (1-Acc)) is the ratio of the number of correctly identified examples to the total number of test examples that are returned by a learning system. Mathematically, accuracy (Acc) is defined as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.24}$$

Although, it is assumed that Acc is a proper measure to evaluate the performance of a predictor, but sometimes it is unsuccessful to compute the real performance of the system. For instance, in the presence of imbalanced data, usually, predictor (or developed model) biased towards the majority class and thereby the higher accuracy achieved by the predictor which does not reflect the real performance of the entire system.

2.6.2 Sensitivity

Sensitivity (Sn) or recall measures the proportion of positives, which are correctly identified by the classifier, i.e., the true positive rate of the prediction system. Numerically, sensitivity is the number of true positive results divided by the sum of true positive and false negative results. This measure is computed as:

$$Sn = \frac{TP}{TP + FN} \tag{2.25}$$

2.6.3 Specificity

Specificity (Sp) measures the proportion of negatives, which are correctly identified by the classifier, i.e., the true negative rate of the prediction system. Mathematically,

specificity is the number of true negative results divided by the sum of true negative and false positive results, i.e.

$$Sp = \frac{TN}{TN + FP} \tag{2.26}$$

Both Sn and Sp have important role in the computation of Acc. The value of Acc measure is influenced by the higher value of Sn or Sp measure. For instance, if we have high Sn and low Sp then Acc becomes biased towards Sn and vice versa is also true. High values of Sn and Sp produce high Acc.

2.6.4 G-Mean

G-mean (G_{mean}) is a geometric mean between Sn and Sp. A high value of G_{mean} indicates a good value of Sn and Sp.

$$G_{mean} = \sqrt{Sn * Sp} \tag{2.27}$$

2.6.5 F-Score

F-Score (F_{Score}) is the harmonic mean of precision (Prc) and Sn of the test. Precision is the ratio of TP to the number of predicted positives. It shows the amount of predicted cancer proteins that are actually related to cancer.

$$Prc = \frac{TP}{TP + FP} \tag{2.28}$$

Prc is used to compute the F_{Score} i.e.,

$$F_{Score} = 2\frac{Prc * Sn}{(Prc + Sn)} \tag{2.29}$$

A high value of F_{Score} shows a high value of both Prc and Sn. It is utilized in tasks where the prediction system is needed to correctly predict examples of a specific class without predicting numerous examples of other classes.

2.6.6 Mathews Correlation Coefficient

Mathews Correlation Coefficient (MCC) is a statistical measure employed to assess the quality of learning algorithms. To express confusion matrix perfectly by a single number, MCC is considered one of the best measures because other measures such as Acc are not useful when the dataset is imbalance. MCC returns values in the range [-1, 1], where 1 represents a perfect prediction, 0 an average random prediction and −1 an inverse prediction.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}} \tag{2.30}$$

2.6.7 Receiver Operating Characteristics

ROC curves are used for analyzing the prediction performance [90]. The information of ROC curves is helpful in the selection of appropriate predictor under the certain decision criterion. The improvement in ROC curves represents low values of false positive rate (1–Sp) and high values of true positive rate (Sn). These values help the points shifting towards the upper left corner of the ROC and thus providing better decision. This kind of behavior is desirable in those applications where the cost of FPR is too important. For example, a weak patient cannot afford high FPR. Minor damage of healthy tissues may be a matter of life and death. On the other hand, when attempts are made to reduce FPR by simply adjusting the decision threshold, the risk of false negative cases might rise in a poor prediction model. This kind of prediction model, specifically in medical applications, might cause the high misclassification cost in various fatal diseases such as lungs, liver, and breast/colon cancer. The values of Area Under Convex Hull (AUCH) are calculated for assessing the performance of the predictor and the examination of predictor consistency.

2.6.8 The Q - Statistic

Yule's Q-statistic is used to measure the diversity among the member predictors in an ensemble. This performance measure of diversity will enhance the prediction performance of the proposed ensemble system. This statistic is used for two base predictors $(C_i \text{ and } C_j)$ as $Q_{i,j} = \frac{ad - bc}{ad + bc}$, where a and d represent the scores of $C_i \text{ and } C_j$ predictors for correct and incorrect predictions, respectively. Whereas, b shows the score when C_i predictor is correct and C_j is incorrect; c is the score of C_j predictor being correct and C_i incorrect. This measure is related to the distance measure that finds the normalized difference between the agreement and disagreement of the two predictors. Q values lie in the range [-1, 1]. The two predictors are statistically independent, if the value of Q is closer to zero. A positive value of Q indicates two predictors tend to agree on the same decisions. However, its negative

value specifies two predictors tend to commit errors on different objects. To generalize the pairwise diversity measure to an entire ensemble which consisting of multiple predictors, we took average Q values of all pair of individual predictors. For ensemble system of *m* base predictors, the averaged Q over all pairs of predictors is computed using Equation 2.31.

$$Q_{avg} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} Q_{i,k} \tag{2.31}$$

2.6.9 Relative Improvement

Relative improvement (RIA) measure is performed to evaluate the relative improvement in approaches. It is defined as:

$$RIA = \sum \frac{\alpha_l - \alpha_i'}{\alpha_i'} \tag{2.32}$$

where α_l represents the performance of the proposed approach in the *i*th dataset and α_i' denotes the performance of the approach being compared with the proposed approach.

In the next chapter, individual cancer prediction systems using KNN and SVM approaches are developed using imbalanced datasets of protein amino acid sequences.

Chapter 3: Individual Prediction Systems

This chapter discusses novel individual prediction systems capable of handling efficiently imbalanced data of protein amino acid sequences. In imbalanced dataset, the number of examples of negative class is appreciably higher than the number of examples from positive class. It is established that the imbalanced data could provide the biased accuracy of the classifiers if they are not designed to make specific arrangement for class imbalance. For this intention, oversampling based MTD technique is employed in feature space to produce diffuse samples of minority class. These new features are presented to MTD-KNN and MTD-SVM prediction systems for better performance. Accuracy, sensitivity, specificity, G-mean, F-Score, MCC, and ROC are employed as performance measures to assess the quality of the proposed prediction systems.

3.1 The Proposed Individual System

The block diagram of the proposed individual system in the presence of imbalanced data for the prediction of human protein related to breast/colon cancer is shown in Fig. 3.1. This figure shows the preprocessing phase and the model development phase.

3.1.1 Preprocessing Phase

In preprocessing phase, primary protein sequences are transformed in AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces. Since, the size of medical data of positive examples is small as compared to negative examples. Therefore, during model development, the decision boundary established by classifier intends to be biased towards majority class. As a result, the performance related to the minority class is also affected. This type of problem is managed by applying different data balancing approaches to create synthetic samples for the minority class. In SMOTE, ‘‘synthetic’’ samples are created instead of simply duplicating [91]. SMOTE method has proved success in the field of handwritten character recognition [92]. For small data set analysis, the MTD technique was proposed in [93]. It was used as the

minority class distribution to generate ‘‘diffuse” samples. This technique is used successfully to balance the medical data.

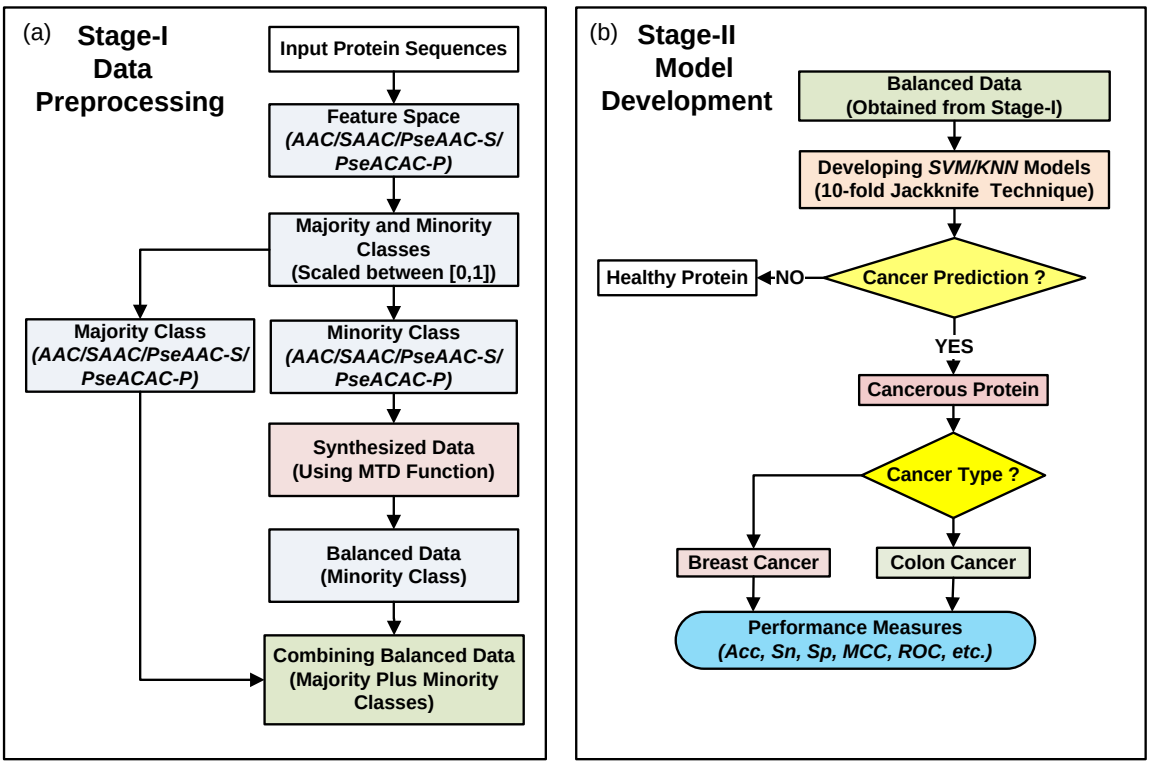


Figure 3.1 Block diagram of the proposed individual system for prediction of protein related to breast/colon cancer. (a) Phase-I represents the preprocessing steps and (b) Phase-II indicates the model development approach.

3.1.2 Imbalanced Data Problem

The issue of class imbalance is investigated by researchers and it is still an active area of research [94-96]. The problem of imbalanced datasets is primarily crucial when the goal is to maximize recognition of the minority class. The imbalanced dataset resulted in suboptimal performance of traditional KNN, NB, and SVM algorithms. Such traditional types of algorithms could be biased toward majority class due to over-prevalence. For example, in case of KNN, the nearest-neighbors are frequently from the majority class and in NB majority class has highest probable than minority class. For SVM, a large number of support vectors are produced from minority class cased overfitting and consequently performance biased towards the majority class. The minority class is customarily the class of interest such as diagnostic medical data and the errors coming from this class is more important. In order to reduce the minority class error, it is worthwhile to develop a system that can make the correct prediction on the minority class effectively and consequently ameliorate the true performance.

The solutions offered to data imbalance are roughly divided into two types. In the first type, preprocessing of input data is carried out to establish class balance. This can be dealt by upsizing the minority class [91] or downsizing the majority class [97]. In the second type, learning algorithm is modified to cope with imbalanced data. This can be handled by building cost-sensitive classifiers that assign a higher cost to misclassification of minority class members [98].

MTD Technique

As described in the section 3.1.2, oversampling and/or under-sampling techniques can be utilized to handle the imbalanced problem. However, under-sampling can discard useful medical and biological information of the majority data class that could be important for the induction process. For example, given imbalance ratio of 100:5, in order to get a close match for the minority class, it might be undesirable to throw away 95% of majority class instances. Therefore, to avoid the risk of deleting useful information of the majority data class, MTD function for upsizing the minority class in feature space, i.e., to generate diffuse samples of the minority class is adopted. The MTD approach uses a membership function, instead of normal distribution assumption, to calculate the possibility values of synthetic [99]. The detailed steps to construct MTD function are given below.

Consider, the given sample $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_n\}$ of minority class, the boundaries *L* and *U* are computed as follows:

$$L = u_{set} - skew_L \sqrt{(-2)\hat{s}_x^2 / N_L \ln(\varphi(L))} \tag{3.1}$$

$$U = u_{set} + skew_U \sqrt{(-2)\hat{s}_x^2 / N_U \ln(\varphi(U))} \tag{3.2}$$

where, $u_{set} = (\min + \max)/2$, and $\hat{s}_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n(n-1)}$ is the variance of *X*, N_L and N_U are the number of data points smaller and greater than u_{set}, respectively, $skew_L = N_L/(N_L + N_U)$ and $skew_U = N_U/(N_L + N_U)$ characterize the degree of asymmetry of a distribution around the dataset core. Here, $\varphi(t)$ is a very small real number (10^{-20}).

In Fig. 3.2, the possibility value of x is assigned by the value of the membership function $M(x)$. The *MTD* function is thus defined as:

$$M(x)=\begin{cases}(x-L)/(u_{set}-L) & \text{for} \quad L\le x\le u_{set}\\ (U-x)/(U-u_{set}) & \text{for} \quad u_{set}\le x\le U\\ 0 & Otherwise\end{cases} \tag{3.3}$$

In function $M(x)$, within the range L and U, diffused samples are generated randomly and MTD values represent the degree of possibility. Here, the values of $\varphi(L)$ and $\varphi(U)$ are set equal to a very small real number (10^{-20}). Because, the degree of possibility of $M(x)$ for two given dataset range limits L and U should tend to zero.

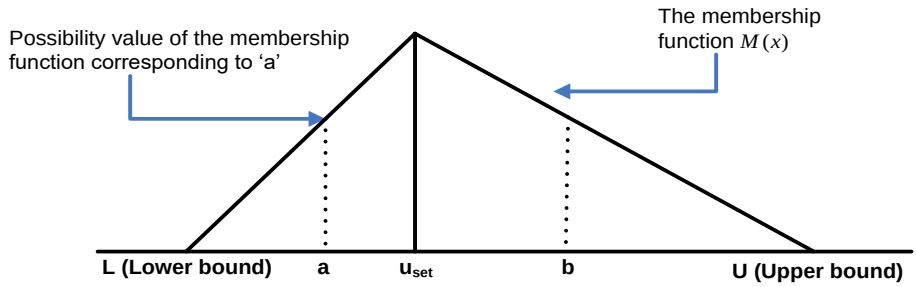


Figure 3.2 MTD technique utilized as membership functions of protein features.

For data balancing, data of minority classes of cancer, breast-cancer, and colon-cancer are generated for each feature space. MTD function is applied to increase the minority class samples of three datasets by the factor of imbalance ratio that is 4.53, 7.66, and 14.30 for datasets of C/NC, B/NBC, and CC/NCC respectively. The detailed description of three datasets before and after applying the MTD technique is demonstrated in Tables 3.1-3.3. This is carried out by randomly generating a value between the previously computed boundaries of L and U (Equations 3.1 and 3.2) and only the values that have a high membership value $M(x)$ are kept, otherwise a new value will be generated. This will ultimately provide us with data, which is close to the real condition of a protein feature space data and in addition, narrow the imbalance gap. The strengths of this method being that we do not have to strictly follow with the assumption that the protein feature space data is normally distributed and in addition, the new data generated is relatively close to the true distribution of the original data. In this context, experiments are conducted to observe the statistical distribution of synthesized data matches with that of the real data. Fig. 3.3 highlights the frequency distributions of breast cancer dataset (a) before and (b) after applying MTD for breaast cancer AAC feature space. However, the proposed technique has the limitation of synthetic data generation that no one to over amplify the dataset as this may inadvertently cause a reverse imbalance of the classes.

Table 3.1 Dataset of C/NC related-protein sequence before and after using MTD.

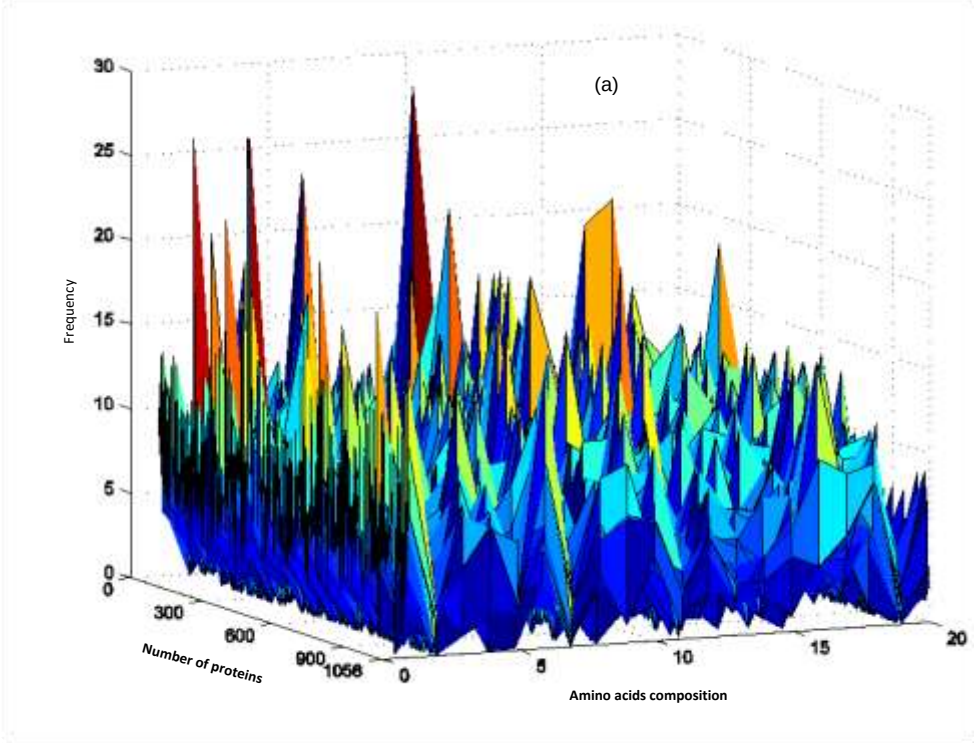
Feature space	Number of C		Number of NC		Total Number of C and NC	
	Before MTD	After MTD	Before MTD	After MTD	Before MTD	After MTD
AAC	191	865	865	865	1056	1730
SAAC	191	865	865	865	1056	1730
PseAAC-S	191	865	865	865	1056	1730
PseAAC-P	191	865	865	865	1056	1730

Table 3.2 Dataset of B/NBC related-protein sequence before and after using MTD.

Feature space	Number of BC		Number of NBC		Total Number of BC and NBC	
	Before MTD	After MTD	Before MTD	After MTD	Before MTD	After MTD
AAC	122	934	934	934	1056	1868
SAAC	122	934	934	934	1056	1868
PseAAC-S	122	934	934	934	1056	1868
PseAAC-P	122	934	934	934	1056	1868

Table 3.3 Dataset of CC/NCC related-protein sequence before and after using MTD.

Feature space	Number of CC		Number of NCC		Total Number of CC and NCC	
	Before MTD	After MTD	Before MTD	After MTD	Before MTD	After MTD
AAC	69	987	987	987	1056	1974
SAAC	69	987	987	987	1056	1974
PseAAC-S	69	987	987	987	1056	1974
PseAAC-P	69	987	987	987	1056	1974



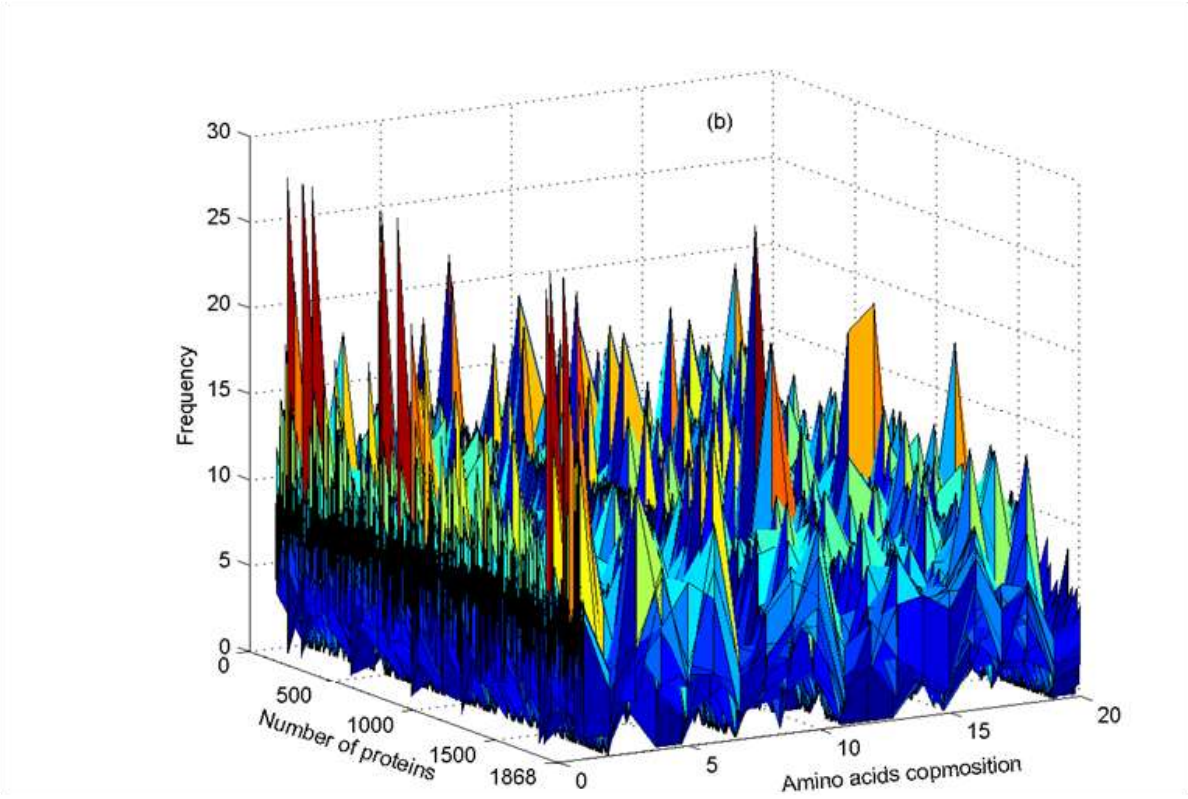


Figure 3.3 Frequency distributions of breast cancer dataset (a) before and (b) after applying MTD for AAC feature space.

3.1.3 Model Development

In model development phase, the balanced data obtained from previous phase is used as input to develop KNN and SVM models. However, for comparison, imbalanced data is used as well. Various data sampling techniques of holdout, boosting, leave one out, and jackknife (cross-validation) are often employed for model development and to assess the performance [100]. The jackknife technique is considered the most rigorous; owing to its ability of yielding unique results. In this thesis, the 10-fold jackknife technique is employed because it is commonly used to examine the performance of predictors such as KNN and SVM [20, 101, 102]. In tenfold Jackknife technique, dataset is divided into ten parts. The 9/10th parts of the dataset are used to train the model and the remaining 1/10th part of the dataset was utilized to test the model. This step is repeated ten times using different training/testing data and the average performance of the model is reported.

3.2 Results and Discussion

The performance of the proposed system is investigated in terms of Acc, Sn, Sp, G_{mean}, F_{Score}, and MCC using: (i) C/NC, (ii) B/NBC, and (iii) CC/NCC protein datasets. Experiments are conducted to explore the effectiveness of various feature spaces AAC, SAAC, PseAAC-S, and PseAAC-P. Experimental results without and with MTD function is examined. In the following subsections, the prediction performance of KNN and SVM models is discussed. Furthermore, a comparative analysis is carried out with NB and QPDR models.

3.2.1 Performance of KNN Models

KNN predictor is tuned by varying the values of *K* for different feature spaces. Those values of *K* are selected which gave minimum prediction errors. Fig. 3.4a and Fig. 3.4b depict prediction error of KNN predictors against the number of nearest neighbors for C/NC and B/NBC datasets, respectively. In these figures, the number of neighbors (*K*) is approximately evenly spaced on a logarithmic scale. These figures helped to select the best nearest neighbors in different spaces.

KNN models are constructed using different balance/imbalance feature spaces for prediction of C/NC, B/NBC, and CC/NCC. Tables 3.4-3.6 show success rates of prediction using KNN_{AAC}, KNN_{SAAC}, $KNN_{PseAAC\text{-}S}$, and $KNN_{PseAAC\text{-}P}$ models for

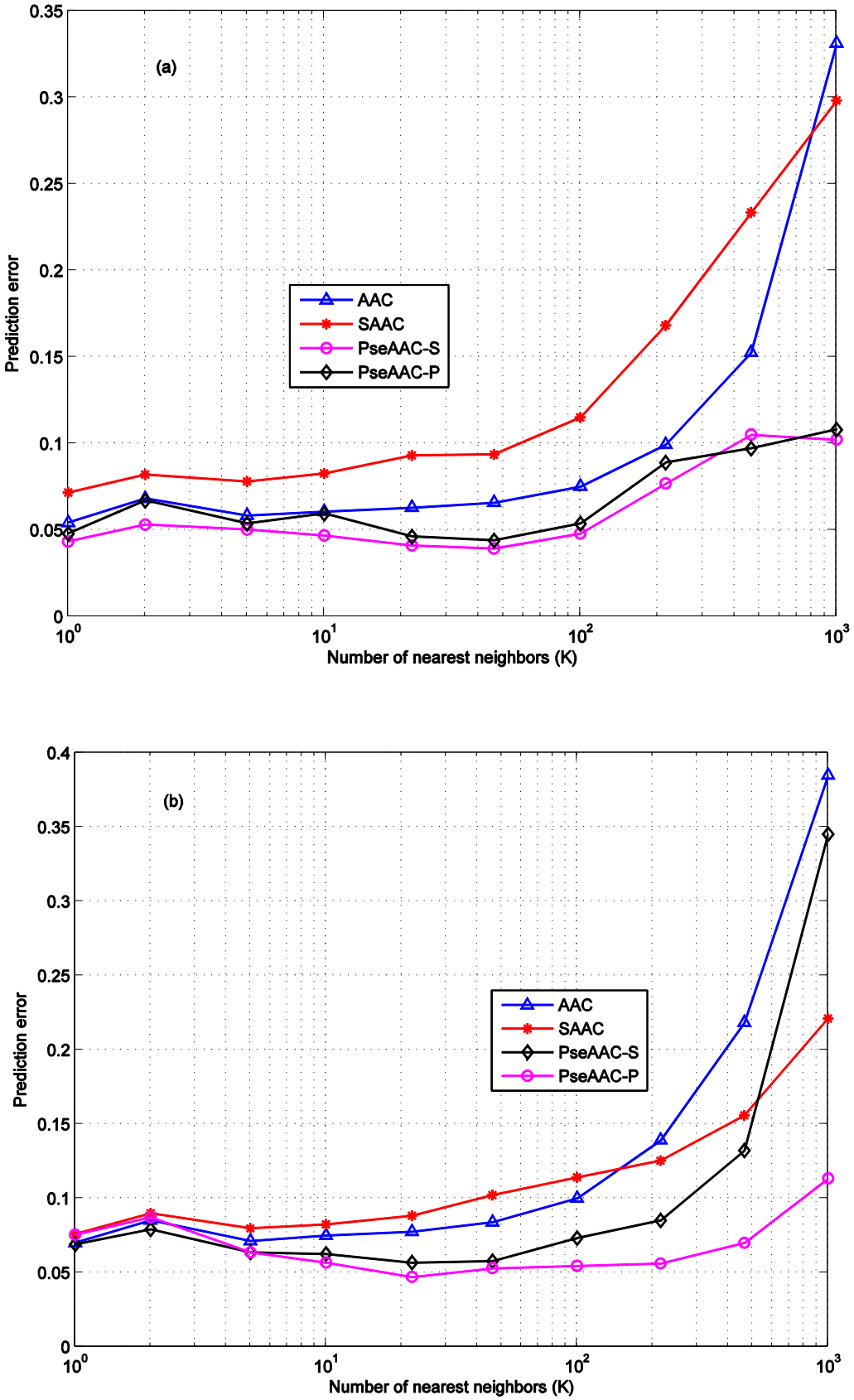


Figure 3.4 Prediction error of KNN predictor vs. number of nearest neighbors for (a) C/NC and (b) B/NBC datasets.

C/NC, B/NBC, and CC/NCC. From Tables 3.4-3.6, it is observed that accuracies for KNN_{AAC}, KNN_{SAAC}, $KNN_{PseAAC\text{-}S}$, and $KNN_{PseAAC\text{-}P}$ models lie in the

Table 3.4 Performance of KNN models for C/NC dataset without/with MTD technique.

Dataset	Model KNN	Acc	Sn	Sp	G_{mean}	F_{score}	MCC
Original data	KNN_{AAC}	88.26	89.51	78.69	83.92	93.10	42.93
	KNN_{SAAC}	88.64	90.47	74.59	82.15	93.37	44.05
	$KNN_{PseAAC\text{-}S}$	88.64	88.87	86.89	87.87	93.26	43.96
	$KNN_{PseAAC\text{-}P}$	87.78	88.01	86.07	87.03	92.72	41.80
Balanced data	KNN_{AAC}	94.74	93.29	96.18	94.73	94.66	63.18
	KNN_{SAAC}	93.99	94.34	93.64	93.99	94.01	64.62
	$KNN_{PseAAC\text{-}S}$	96.01	95.14	96.88	96.01	95.98	65.28
	$KNN_{PseAAC\text{-}P}$	95.49	93.87	97.11	95.48	95.42	63.77

Table 3.5 Performance of KNN models for B/NBC dataset without/with MTD technique.

Dataset	Model KNN	Acc	Sn	Sp	G_{mean}	F_{score}	MCC
Original data	KNN_{AAC}	89.02	94.65	45.90	65.91	93.84	46.52
	KNN_{SAAC}	90.15	97.75	31.97	55.90	94.61	59.85
	$KNN_{PseAAC\text{-}S}$	90.81	94.97	59.02	74.86	94.82	54.62
	$KNN_{PseAAC\text{-}P}$	89.58	94.33	53.28	70.89	94.12	48.90
Balanced data	KNN_{AAC}	93.36	93.58	93.15	93.36	93.38	63.74
	KNN_{SAAC}	92.56	92.61	92.51	92.56	92.56	62.63
	$KNN_{PseAAC\text{-}S}$	94.54	94.43	94.65	94.54	94.53	64.64
	$KNN_{PseAAC\text{-}P}$	94.59	94.33	94.86	94.59	94.58	64.50

Table 3.6 Performance of KNN models for CC/NCC dataset without/with MTD technique.

Dataset	Model KNN	Acc	Sn	Sp	G_{mean}	F_{score}	MCC
Original data	KNN_{AAC}	92.90	98.99	5.80	23.95	96.30	23.65
	KNN_{SAAC}	93.47	99.70	4.35	20.82	96.61	46.50
	$KNN_{PseAAC\text{-}S}$	92.42	98.18	10.14	31.56	96.04	23.10
	$KNN_{PseAAC\text{-}P}$	92.42	98.18	10.14	31.56	96.04	23.10
Balanced data	KNN_{AAC}	95.39	96.96	93.82	95.38	95.46	67.86
	KNN_{SAAC}	93.62	91.79	95.44	93.60	93.50	61.51
	$KNN_{PseAAC\text{-}S}$	96.05	97.97	94.12	96.03	96.12	69.11
	$KNN_{PseAAC\text{-}P}$	95.54	97.26	93.82	95.53	95.62	68.24

range 87.78% - 88.26% for C/NC, 89.02% - 90.81% for B/NBC and 92.42% - 93.47% for CC/NCC without applied MTD technique. KNN_{SAAC} model has given the best accuracy of 93.47% among all other models for prediction of CC/NCC. However, when MTD technique is applied, the accuracy approaches to the highest value of 96.05% using $KNN_{PseAAC\text{-}S}$ model for CC/NCC. The average enhancement in accuracy using KNN models is 6.73%, 3.87%, and 2.35% for the prediction of C/NC, B/NBC, and CC/NCC, respectively. This highlights the usefulness of the proposed approach with data balancing technique.

Tables 3.4-3.6 demonstrate, without balanced data, KNN_{SAAC} models have given the higher Sn measure of 90.47%, 97.75%, and 99.70% for C/NC, B/NBC, and CC/NCC, respectively. In the case of Sp measure, $KNN_{PseAAC\text{-}S}$ models have provided values of 86.89%, 59.02%, and 10.14% for C/NC, B/NBC, and CC/NCC, respectively.

It is noted that, without balanced data, KNN models have higher values of Sn and lower values of Sp. The difference between Sn and Sp depend on the type of cancer and the prediction model. It is observed that, without balanced data, SAAC feature space is more sensitive to Sn and Sp values as compared to other feature spaces. For example, KNN_{SAAC} models have maximum differences of 15.88%, 65.78%, and 95.35% in the values of Sn and Sp for the prediction of C/NC, B/NBC, and CC/NCC, respectively. Particularly, in case of CC/NCC this difference is much higher (95.35%) and over inflates Sp. In fact, the predictor is biased towards majority class samples. As evident from Tables 3.4-3.6, when data balancing technique is introduced, enough improvement is obtained in the values of Sp. The higher values of Sn and Sp are desired for better medical decision. G_{mean} and F_{Score} are also improved because these values depend on Sn and Sp, hence, better performance is obtained using data balancing technique.

Without balanced data, KNN_{SAAC} model has provided the best MCC value of 44.05%, followed by $KNN_{PseAAC\text{-}S}$, (43.96%), KNN_{AAC}, (42.93%), and $KNN_{PseAAC\text{-}P}$, (41.80%) models for cancer and non-cancer prediction. However, without balanced data, KNN has given the best MCC value of 59.85% for B/NBC dataset using SAAC feature space. With balanced data, $KNN_{PseAAC\text{-}S}$ model has obtained the highest MCC value of 69.11% for CC/NCC than other KNN models. However, it is observed that,

on average with balanced data, KNN models are enhanced the MCC values of 21.03%, 11.40%, and 37.59% for C/NC, B/NBC, and CC/NCC, respectively. Thus, it is inferred that data balancing technique has provided adequate improvement in the prediction of KNN models.

3.2.2 Performance of SVM Models

In the training phase of SVM, several experiments were performed by adding diffuse samples in minority class datasets. In our case, dataset of *BC* proteins represents the minority class. *BC* dataset was increased by a step size of 122. The majority class, i.e., *NBC* proteins, was fixed to 934. Fig. 3.5 shows the number of SVs produced during training of SVM with the varying dataset size of the minority class. From Fig. 3.5, it can be observed that the number of SVs is decreasing with increasing the number of proteins (diffuse points) for minority class. The number of SVs greater than 10% of the original dataset would result in overfitting. To avoid overfitting, the size of dataset for minority class that produced approximately 10% the number of SVs is selected.

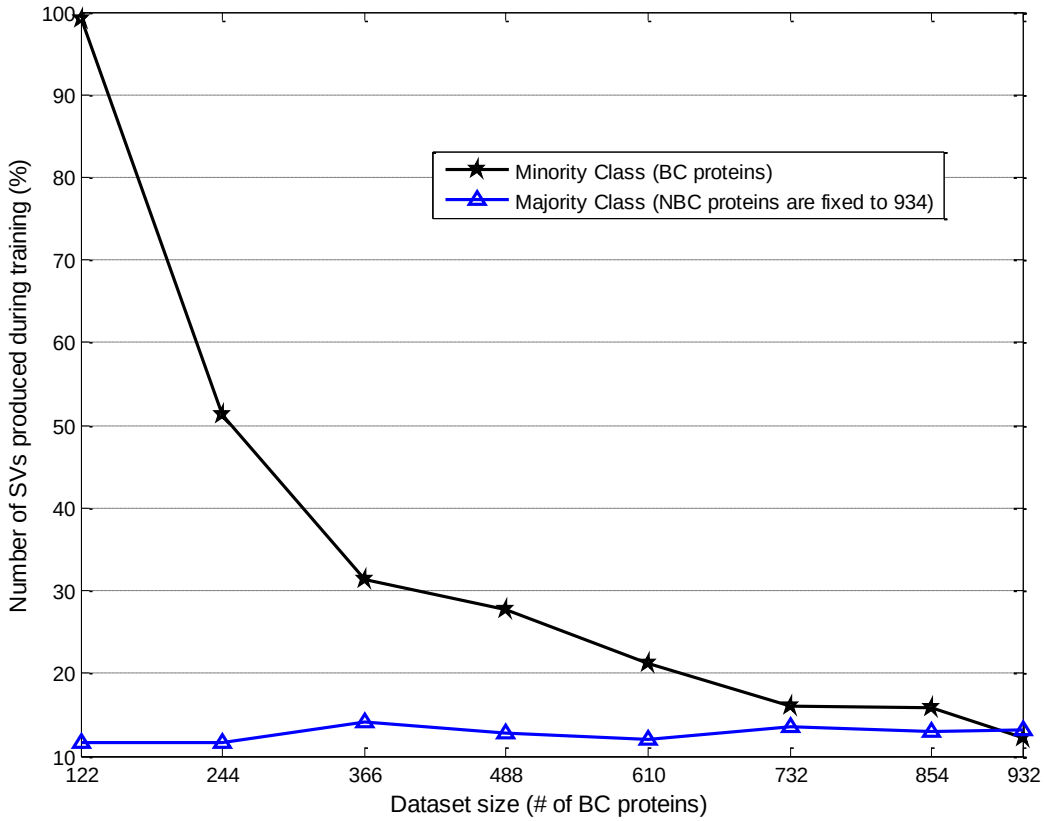


Figure 3.5 In the training phase of SVM, the number of SVs is decreasing with increasing the number of proteins (diffuse points) for minority class dataset. To avoid overfitting, the size of dataset for minority class is chosen that produced approximately 10% the number of SVs.

In order to evaluate the performance of SVM predictor, several experiments were performed by varying the number of samples of minority class. Diffuse samples are added in the minority class datasets because it helps in avoiding overfitting of SVM. Fig. 3.6 illustrates the performance of SVM predictor for breast cancer dataset. From Fig. 3.6, it is observed that Acc and Sp are not dependent on the size of the dataset. However, the performance of SVM, in terms of Sp, G-mean, F-score, and MCC is enhanced up to a certain limit. It is observed that the *Sn* of classification is considerable improved by varying the imbalance ratio of the classes. This result is in accordance with Zubair, et al. [103].

SVM models are constructed using optimal parameter values (C, σ) for various feature spaces AAC, SAAC, PseAAC-S, and PseAAC-P. The best SVM model is selected using the appropriate number of support vectors. Enhanced performance of SVM models is acquired using data balancing technique. The performance of SVM models is reported for: (i) C/NC, (ii) B/NBC, and (iii) CC/NCC.

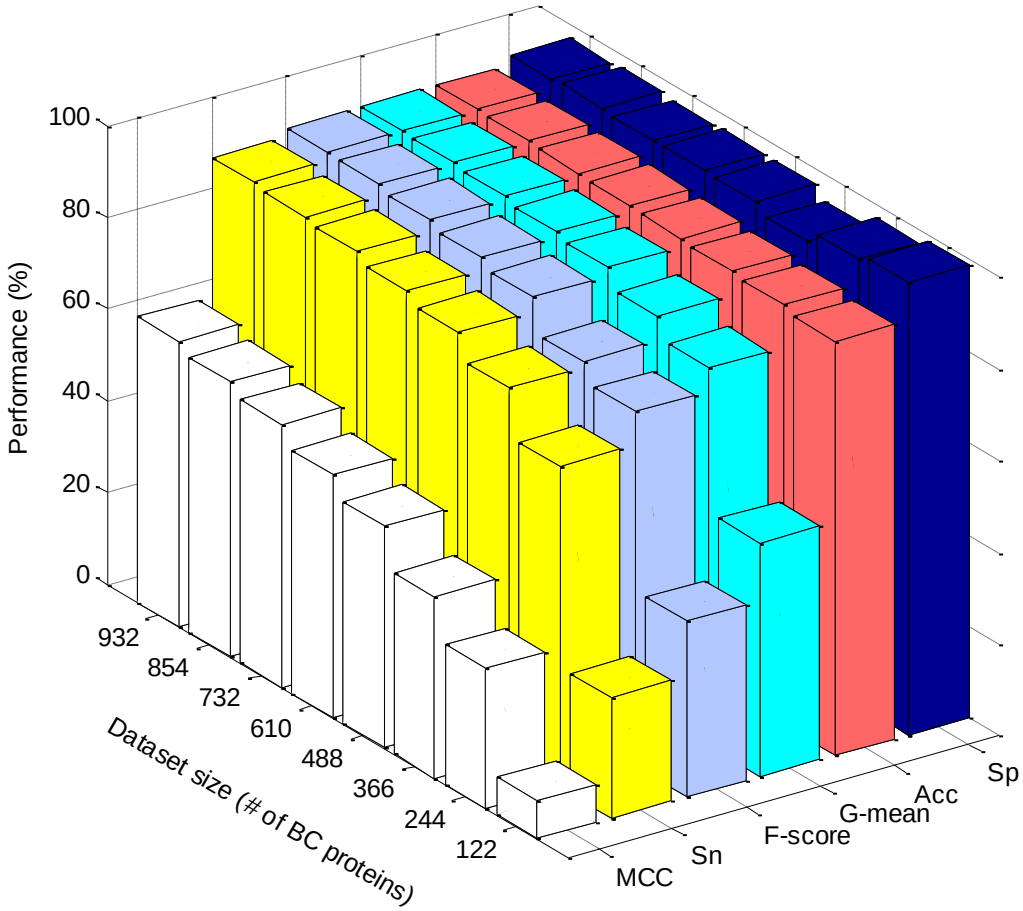


Figure 3.6 Performance of individual-SVM against varying the number of samples of minority class.

Tables 3.7-3.9 demonstrate, with balanced data, the performance of SVM based models for C/NC, B/NBC, and CC/NCC. It is observed from Table 3.7 that among SVM models, $SVM_{PseAAC-S}$ model has the highest accuracy of 96.71% for C/NC. However, for same cancer, other SVM models have given accuracy around 95.97%. It is noted from Table 3.8 that $SVM_{PseAAC-S}$ model provided the best accuracy of 95.18% for B/NBC. However, other SVM models have given accuracy near to 94.88%. In the case of colon cancer (Table 3.9), again, $SVM_{PseAAC-S}$ model has provided the highest accuracy value of 96.50% and remaining SVM models have accuracy about 95.74%. Therefore, it is deduced that $SVM_{PseAAC-S}$ model gives the best accuracy among other feature spaces.

It is observed from Table 3.7-3.9 that SVM model has provided higher Sn values of 97.69%, 93.04%, and 100.00% using PseAAC-P, PseAAC-S, and

Table 3.7 Performance of SVM models for C/NC with balanced data.

Model	C/NC					
MTD-SVM	Acc	Sn	Sp	G_{mean}	F_{score}	MCC
SVM_{AAC}	95.72	96.88	94.57	95.72	95.77	67.65
SVM_{SAAC}	95.72	96.76	94.68	95.72	95.77	67.49
$SVM_{PseAAC-S}$	96.71	97.57	95.84	96.70	96.73	68.35
$SVM_{PseAAC-P}$	96.47	97.69	95.26	96.47	96.52	68.58

Table 3.8 Performance of SVM models for B/NBC with balanced data.

Model	B/NBC					
MTD-SVM	Acc	Sn	Sp	G_{mean}	F_{score}	MCC
SVM_{AAC}	94.59	92.72	96.47	94.57	94.49	62.50
SVM_{SAAC}	95.07	91.97	91.97	95.02	94.92	61.55
$SVM_{PseAAC-S}$	95.18	93.04	97.32	95.16	95.08	62.80
$SVM_{PseAAC-P}$	94.97	91.97	97.97	94.92	94.81	61.57

Table 3.9 Performance of SVM models for CC/NCC with balanced data.

Model	CC/NCC					
MTD-SVM	Acc	Sn	Sp	G_{mean}	F_{score}	MCC
SVM_{AAC}	95.44	96.76	94.12	67.56	95.50	67.56
SVM_{SAAC}	95.64	96.96	94.33	95.63	95.70	67.79
$SVM_{PseAAC-S}$	96.50	100.00	93.01	96.44	96.62	71.98
$SVM_{PseAAC-P}$	96.15	99.09	93.21	96.11	96.26	70.72

PseAAC-S feature spaces for C/NC, B/NBC, and CC/NCC, respectively. However, $SVM_{PseAAC-S}$, $SVM_{PseAAC-P}$, and SVM_{SAAC} models have provided higher Sp values of 95.84%, 97.97%, and 94.33% for the prediction of C/NC, B/NBC, and CC/NCC, respectively.

With balanced data (Tables 3.7-3.9), it is observed that SVM models have achieved higher value of Sp and consequently, G_{mean} and F_{Score} are also improved. However, it is found that, among other SVM models, $SVM_{PseAAC-S}$ model has attained the highest G_{mean} value of 96.70% and F_{Score} value of 96.73% for C/NC. It is observed $SVM_{PseAAC-S}$ model for CC/NCC has given the highest MCC value of 71.98%. It is found $SVM_{PseAAC-S}$ model has yielded the best MCC values of 68.35% and 62.80% for C/NC and B/NBC, respectively. Therefore, it is concluded, in term of MCC, that PseAAC-S feature space with SVM predictor has generated an excellent discriminant space for the prediction of C/NC, B/NBC, and CC/NCC.

The prediction performance of SVM models is also analyzed in terms of ROC curves using AAC, SAAC, PseAAC-S, and PseAAC-P features spaces. The improved ROC curve is helpful in selecting operating point (threshold) of the predictor. ROC curves for SVM models using different feature spaces, without and with balanced data are shown in Figs. 3.7–3.12.

Fig. 3.7 indicates AUC and AUCH measures of SVM models without balanced data for decision of C/NC disease. SVM_{AAC} model has shown small difference in values of AUCH (0.77) and AUC (0.76) as compared to other SVM models. $SVM_{PseAAC-P}$ has given the lowest values of 0.69 and 0.72 for AUC and AUCH, respectively. However, SVM_{SAAC} model has provided the highest values of AUC (0.78)/AUCH (0.80).

Fig. 3.8 indicates AUC and AUCH measures of SVM models with balanced data for decision of C/NC disease. With balanced data, improvement of 14.44%, 12.82%, 17.83% and 21.86% in AUC is observed for AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces, respectively. $SVM_{PseAAC-P}$ model has gained the highest values of AUC (0.91) and AUCH (0.91). These equal values of AUC and AUCH for PseAAC-P feature space represent consistent and better performance among other models. Fig. 3.9 indicates AUC and AUCH measures using SVM models without balanced data for the decision of B/NBC. It is noted that all four SVM models have

consistent performance for this type of cancer. Fig. 3.10 illustrates AUC and AUCH measures using SVM models with balanced data for B/NBC. With balanced data, improvement in AUC values of 3.40%, 2.41%, 0.32%, 1.86%, and 21.86% is observed using AAC, SAAC, PseAAC-S, and PseAAC-P features spaces, respectively. From this figure, it is observed the lowest value of AUC (0.89) and AUCH (0.91) for $SVM_{PseAAC-P}$ and SVM_{SAAC} models, respectively. However, the highest values of AUC (0.90) and AUCH (0.92) are found using $SVM_{PseAAC-S}$ model. Hence, $SVM_{PseAAC-S}$ models demonstrate better performance for B/NBC.

Fig. 3.11 shows AUC and AUCH measures of SVM models without balanced data using AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces for CC/NCC. Again, the $SVM_{PseAAC-S}$ model produced the higher value of AUC (0.81) and AUCH (0.83) for the prediction of colon cancer. Fig. 3.12 shows AUC and AUCH measures of SVM models with balanced data using AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces for CC/NCC. When applied balanced data, values of AUC (AUCH) are increased by 11.88% (11.64%), 21.62% (20.00%), 8.77% (10.17%), and 16.89% (18.04%) of SVM_{AAC}, SVM_{SAAC}, $SVM_{PseAAC-S}$, and $SVM_{PseAAC-P}$ models, respectively. The SVM_{AAC} model has given the relatively small difference in values of AUCH (0.90) and AUC (0.91). This represents the consistent performance of SVM_{AAC} model. Again, the $SVM_{PseAAC-S}$ model has given the higher AUCH value of 0.93. This AUCH value indicates that $SVM_{PseAAC-S}$ model has provided better decision for cancer disease.

3.2.3 Performance Comparison of Different Models

Table 3.10 shows a comparative analysis, in the term of Acc, of the proposed KNN and SVM models with NB and QPDR models [12] for C/NC, B/NBC, and CC/NCC. Form Table 3.10, it is found that SVM based models have better decision than KNN models for cancerous and non-cancerous related protein sequences. Without balanced data, SVM_{SAAC} models have indicated that SAAC feature space is the best discriminant feature among other feature spaces. However, if only KNN models are considered then $KNN_{PseAAC-S}$ models possessed the best discriminant PseAAC-S feature space. The highest accurate values achieved using KNN and SVM models are 96.05% and 96.71%, respectively. The KNN model has given 2.9% higher accuracy than NB predictor. However, SVM has provided 3.6% higher accuracy than NB using

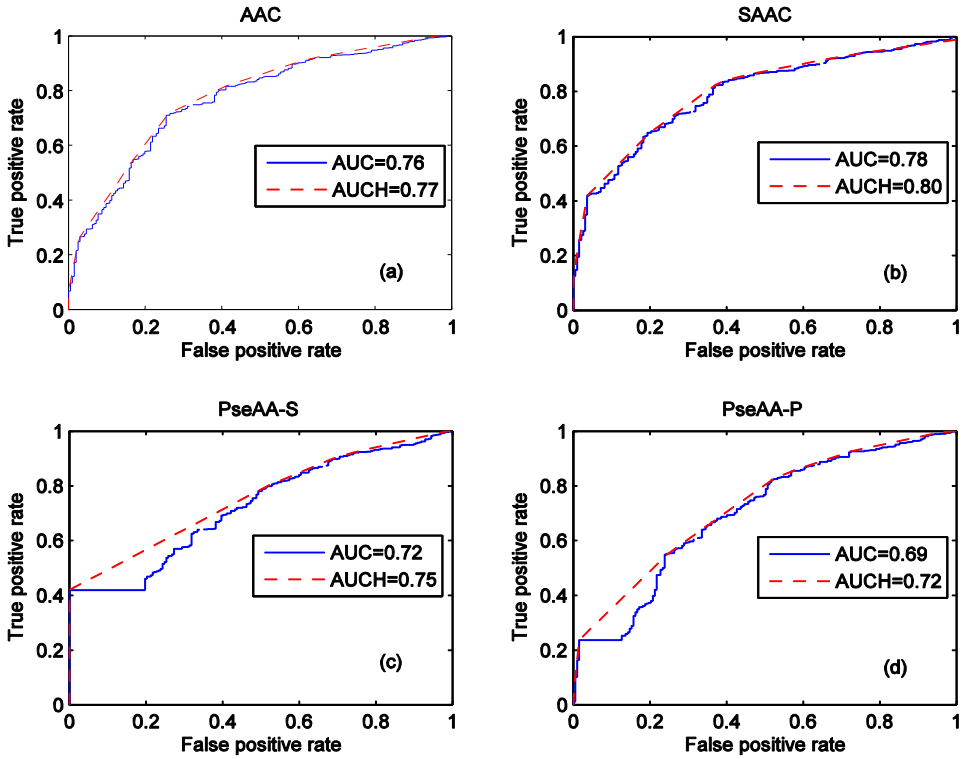


Figure 3.7 ROC curves of SVM models using feature spaces: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P for C/NC without balanced data.

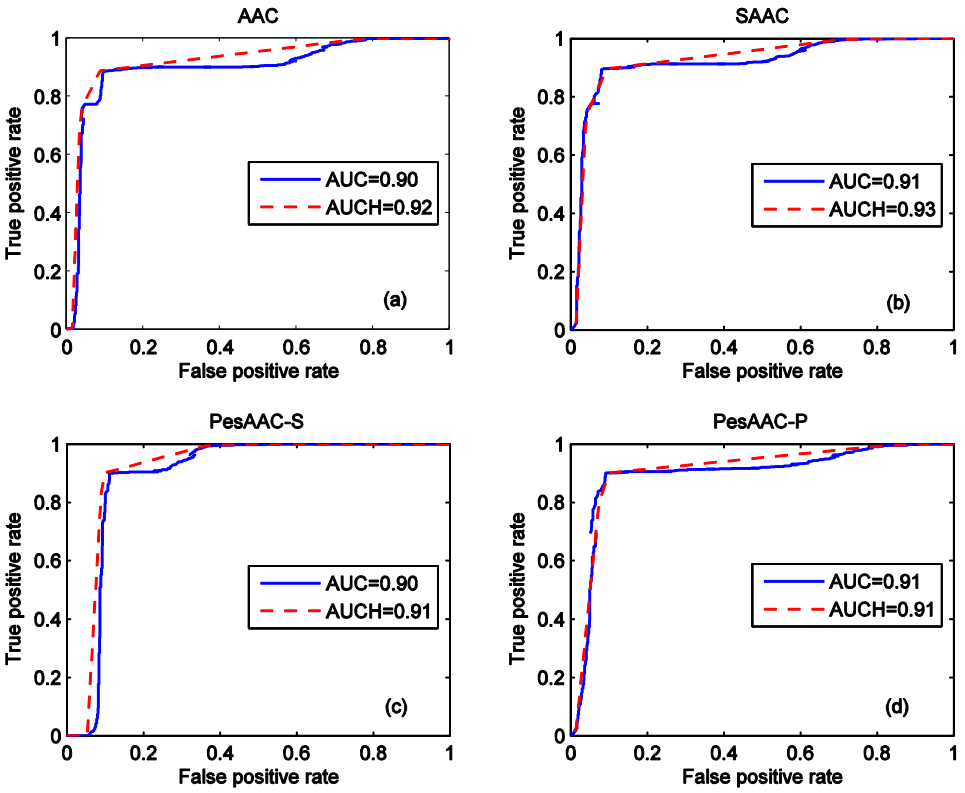


Figure 3.8 ROC curves of SVM models using feature spaces: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P for C/NC disease with balanced data.

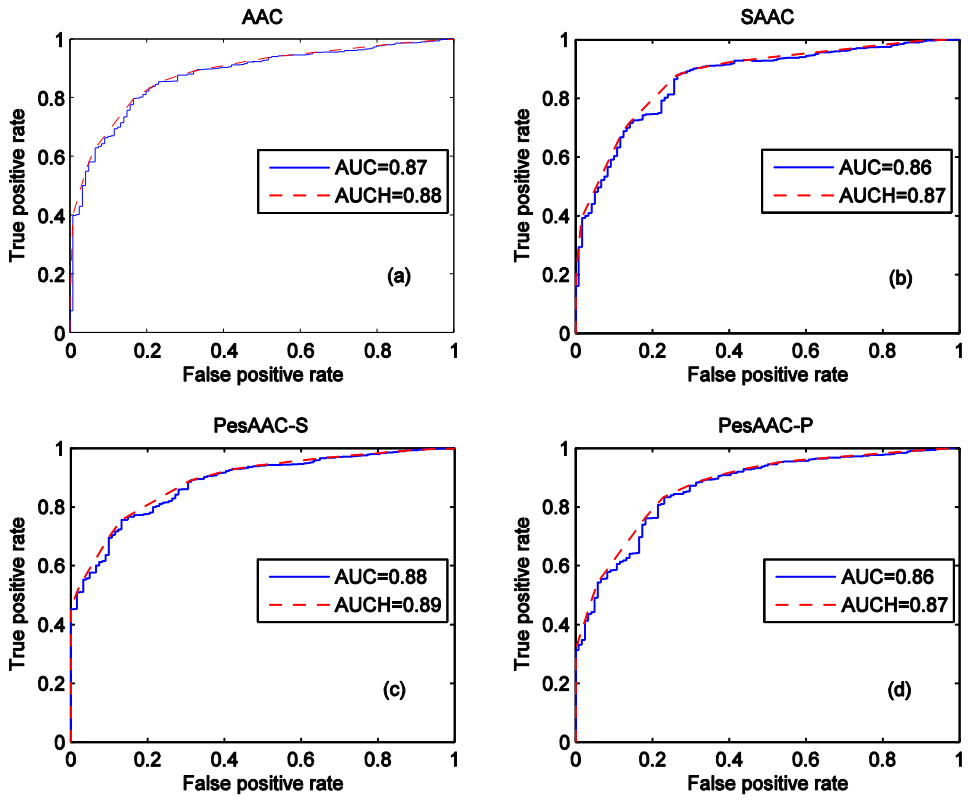


Figure 3.9 ROC curves of SVM models using feature spaces: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P for B/NBC without balanced data.

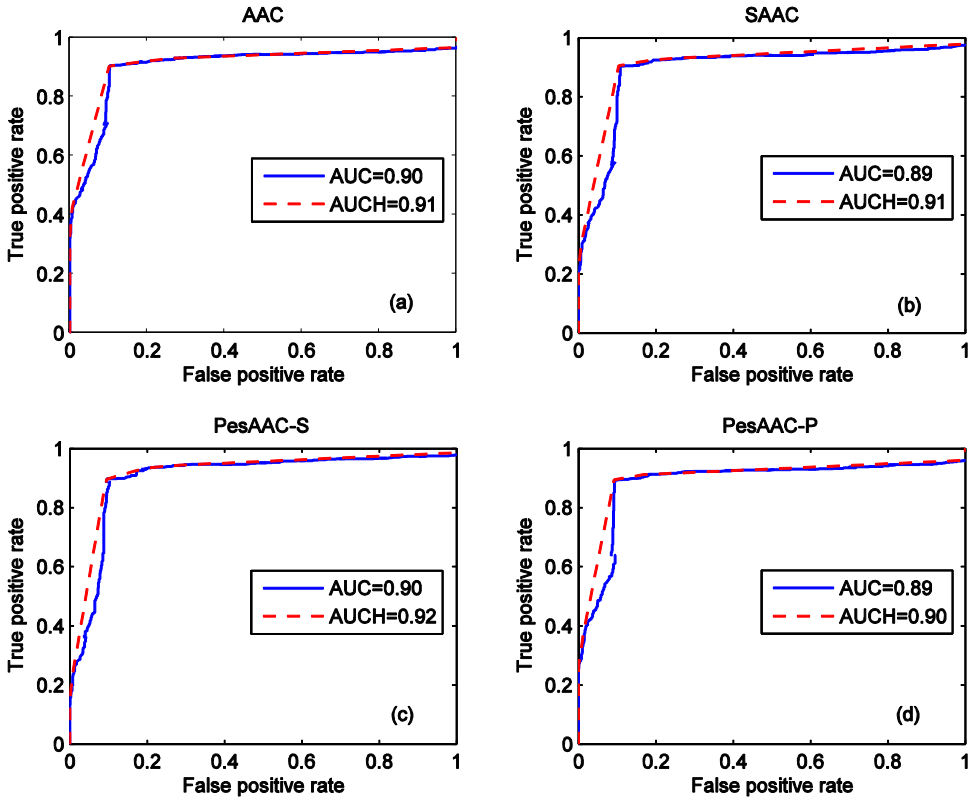


Figure 3.10 ROC curves of SVM models using feature spaces: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P for B/NBC with balanced data.

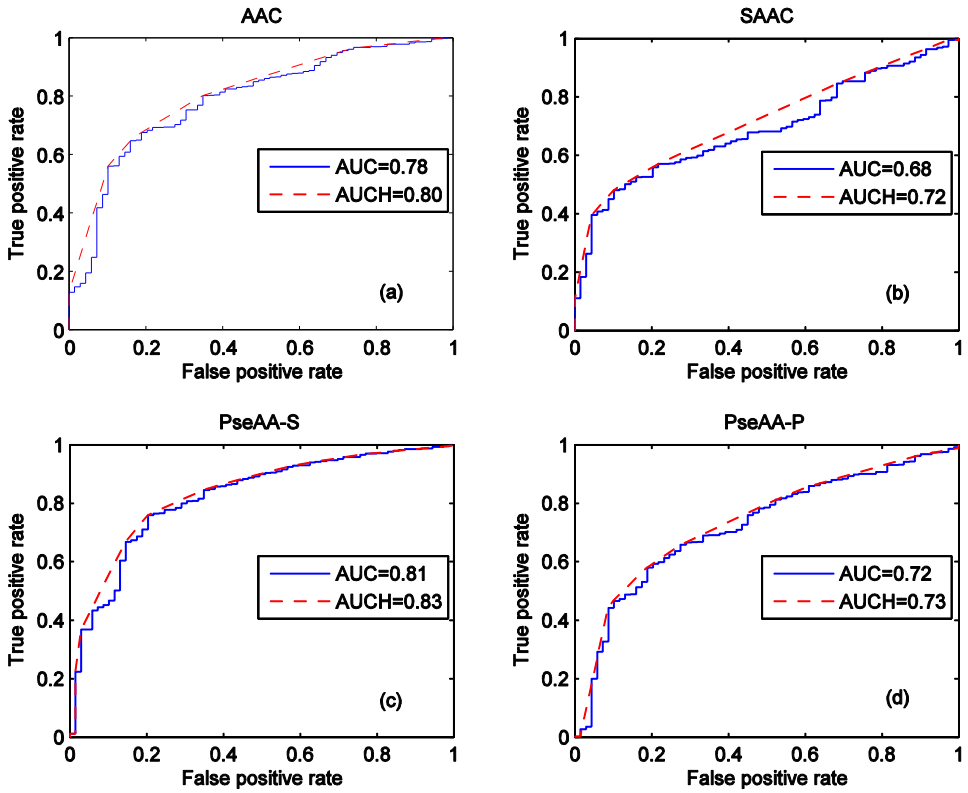


Figure 3.11 ROC curves of SVM models using feature spaces: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P for CC/NCC without balanced data.

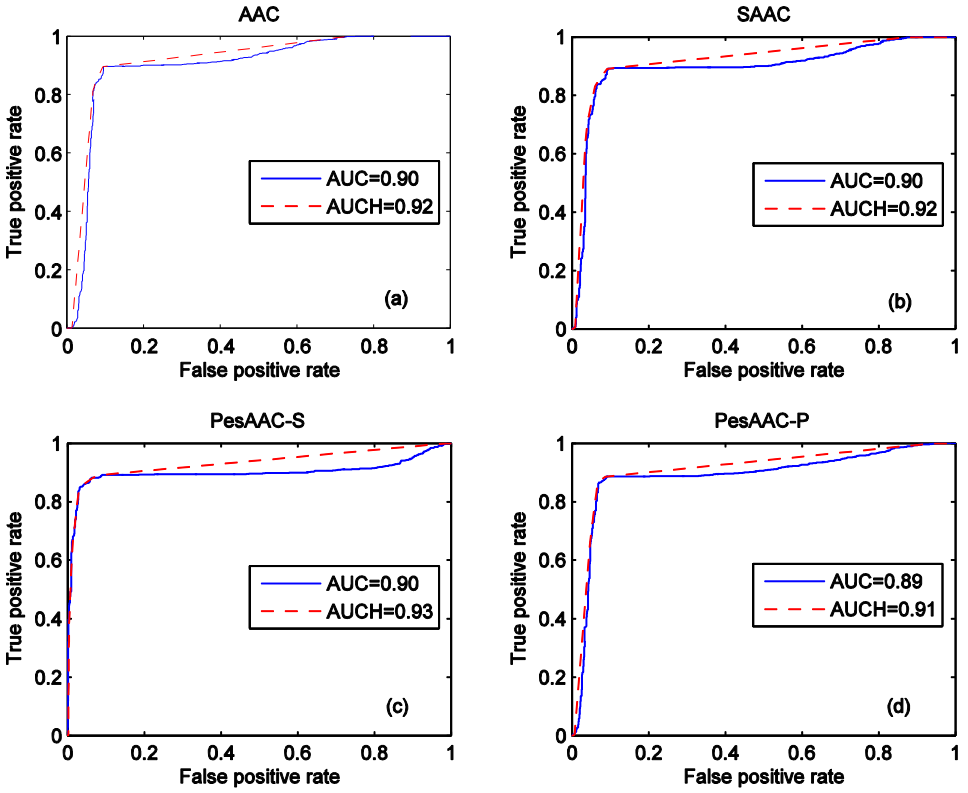


Figure 3.12 ROC curves of SVM models using feature spaces: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P for CC/NCC with balanced data.

C/NC and B/NBC cancer datasets. From Table 3.10, it is observed that all predictors, except KNN, have given accuracy value of 96.50% using different feature spaces of CC/NCC dataset. Therefore, SVM and NB predictors yielded the same level of performance for CC/NCC cancer dataset. For the same protein datasets, QPDR models have reported the highest accuracy value of 91.80%. Therefore, the proposed KNN and SVM models have provided 4.3% and 4.9% higher accuracy over QPDR

Table 3.10 Performance comparison of the proposed prediction models for C/NC, B/NBC, and CC/NCC.

Model	Without balanced data Acc (%)			With balanced data Acc (%)		
	C/NC	B/NBC	CC/NCC	C/NC	B/NBC	CC/NCC
KNN_{AAC}	88.26	89.02	92.90	94.74	93.36	95.39
KNN_{SAAC}	88.64	90.15	93.47	93.99	92.56	93.62
$KNN_{PseAAC-S}$	88.64	90.81	92.42	96.01	94.54	96.05
$KNN_{PseAAC-P}$	87.78	89.58	92.42	95.49	94.59	95.54
SVM_{AAC}	91.21	89.02	93.54	95.72	94.59	95.44
SVM_{SAAC}	93.13	91.01	93.42	95.72	95.07	95.64
$SVM_{PseAAC-S}$	92.01	89.43	93.41	96.71	95.18	96.50
$SVM_{PseAAC-P}$	86.52	89.21	93.30	96.47	94.97	96.15
NB_{AAC}	84.85	82.58	76.42	90.75	93.84	96.35
NB_{SAAC}	88.16	86.65	81.16	88.96	93.47	96.50
$NB_{PseAAC-S}$	80.68	77.56	72.54	91.04	88.06	92.05
$NB_{PseAAC-P}$	81.34	79.45	75.47	93.06	94.16	95.59
$QPDR_{TIe+ dTIe}$[$] [12]	90.00	91.80	88.20	-	-	-
$QPDR_{pTIe}$[$]	90.50	91.80	89.20	-	-	-

[$]dTIe=difference between the same TIs and the average of the TIs for each type of cancer with embedded star graph, pTIe=cancer probability TIs with embedded star graph, for more detail see [12].

models. The Multi-target *QPDR* models provided poor performance because these models were based on multiple linear regression (*MLR*) technique. Due to the linear nature of *MLR* models, it is difficult to model accurately the nonlinearity in the features to corresponding target labels. Therefore, the maximum performance of *QPDR* models is limited to 91.80% for the prediction of cancer. To the best of the author's knowledge, these results are the best-reported results, so far, using the same

datasets. Hence, it is concluded that SVM models outperform over all other predictors for three datasets of cancers.

In this chapter, the improved performance of the individual prediction systems for breast and colon cancers is presented. The analysis indicated that SVM based models are better as compared to KNN, NB, and QPDR models. It has been shown that the presence of balanced data caused the predictors to reduce bias towards the majority class and thereby the overall performance is ameliorated. In the next chapter, the performance of the random ensemble system and cost-sensitive learning will discuss.

Chapter 4: Random Ensemble System

The advancement in the machine learning and pattern recognition resources encouraged researchers to develop efficient ensemble system for the prediction of cancer. These decision making systems utilized different classification algorithms. Most of the original algorithms engage in to minimize the error rate of incorrect prediction of class labels. In the development of unbiased ensemble system, it is important to understand how class imbalance affects data balancing and cost-sensitive learning techniques. In this chapter, we investigate the behavior of ensemble systems by employing data balancing or cost-sensitive learning techniques. For this purpose, two new homogeneous ensemble systems are discussed and their performance is analyzed.

4.1 The Proposed Ensemble Systems

Fig. 4.1 shows the detailed diagram of the proposed homogeneous systems: CanPro-IDSS and Can-CSCGnB. The CanPro-IDSS (Cancer Protein Intelligent Decision Support System) is a random ensemble system developed by combining diffuse minority over-sampling based MTD technique with RF ensemble. A web-server based ready-to-use cancer predictive random ensemble system is developed. This system is publicly accessible at *http://115.167.2.211/canpro-idss/*. Whereas, the Can-CSCGnB (Cancer Prediction Cost-Sensitive Classifier GentleBoost system) is a gentle ensemble system developed by employing CSL technique with GentleBoost ensemble. The proposed systems have following main modules: (i) Input protein dataset, (ii) Feature generation, (iii) Data balancing in Fig. 4.1a, (iv) Formation of training and testing datasets, (v) Development of ensemble models, and (vi) Performance measures.

In the first stage, useful feature information is extracted from the protein sequences using physicochemical properties of Hd and Hb of amino acids. Each feature space between [0, 1] is scaled so that features with large numeric values would not dominate small numeric values. In data balancing module (Fig. 4.1a), MTD approach is utilized. In classifier development stage, ensemble models in individual

feature spaces are developed. However, in Fig. 4.1b, in model development stage, CSL with GentleBoost, AdaBoostM1, and Bagging ensembles is employed.

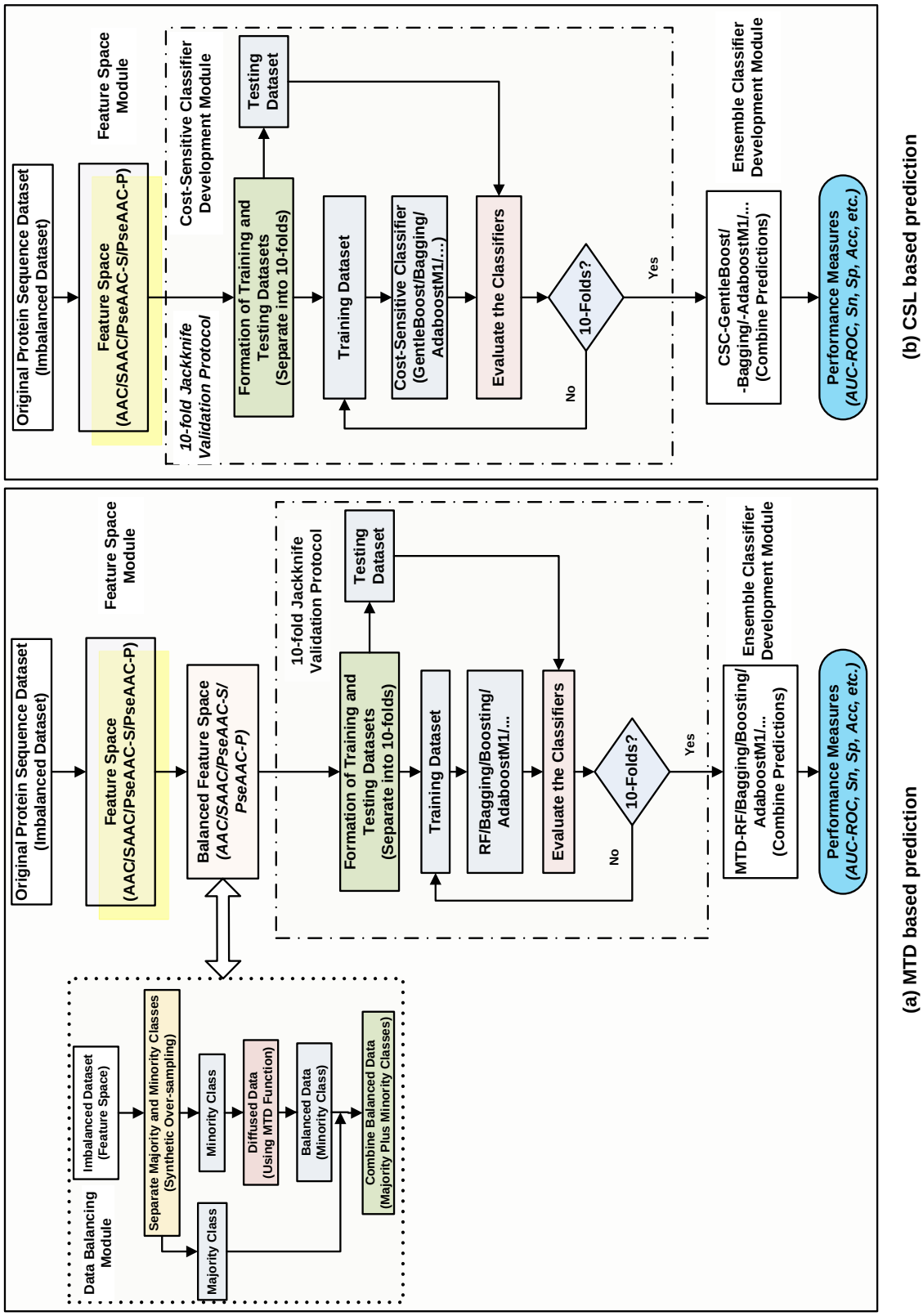


Figure 4.1 Detailed block diagram of the proposed CanPro-IDSS and Can-CSCGnB systems using (a) MTD and (b) CSL techniques, respectively.

4.1.1 Cost-Sensitive Learning Technique

This section is targeted to explore an appropriate technique that is capable to manage imbalanced dataset. Generally, the conventional computational intelligent approaches perform well on balanced datasets. However, if the dataset is imbalanced then it is biased toward the majority class that causes to inflate the sensitivity of the predictions. To account for this problem, the effect of MTD and CSL techniques is explored. MTD function oversamples the minority class in feature space, i.e., to generate diffuse samples of minority class. Alternatively, CSL technique handles with the misclassification costs of the classifier. The detail of MTD technique is discussed in chapter 3. However, CSL, technique will be exploited.

CSL technique is employed to handle imbalanced data problem by assuming misclassification costs. In this research, cost-insensitive algorithm is utilized as cost-sensitive without modifying actual training algorithm. The aim is to develop a model that has the least misclassification costs. Table 4.1 shows the cost matrix for C/NC problem.

Table 4.1 Cost matrix for binary problem.

		Predicted class	
		Minority	Majority
True class	Minority	$C_t(1,1)$	$C_t(1,0)$
	Majority	$C_t(0,1)$	$C_t(0,0)$

In Table 4.1, '1'represents minority class and '0' represents majority class, and C_t(i, j) indicates the cost of misclassifying of example from true class *i* as predicted class *j*.

According to the minimum expected cost principle, a classifier must classify an example $\mathbf{x}$ into class *i*, which has the minimum expected cost. For a given cost matrix, this expected cost $R(i/\mathbf{x})$ is given as

$$R(i/\mathbf{x}) = \sum_j P(j/\mathbf{x})C(i,j) \tag{4.1}$$

where $P(j/\mathbf{x})$ is the probability estimation of classifying an example into class *j*.

4.1.2 Web Server (CanPro-IDSS)

A web based "CanPro-IDSS" system is developed. The proposed CanPro-IDSS

system could be employed by academia, practitioners, or clinician, for the early diagnosis of breast cancer using protein sequences of the affected part of the organ. Protein sequences, which may be taken from DNA sequences, etc., could easily be supplied/presented to the CanPro-IDSS system. If the specific order of amino acids in protein sequence is altered, the system will indicate/diagnose cancer or else non-cancer. When cancer is diagnosed, the practitioner may carry on further by proposing additional standard procedures to examine the severity of the disease. Following procedure explains the use of CanPro-IDSS web server:

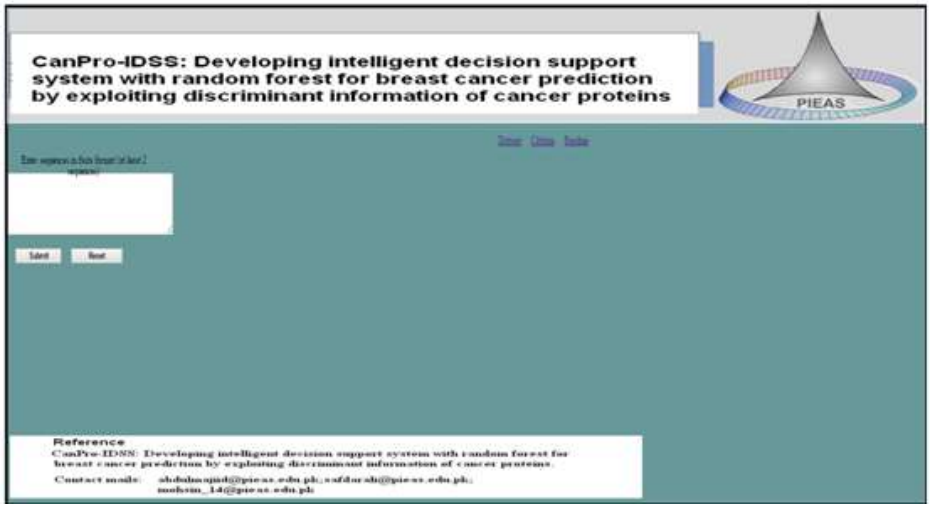
I. Open the web page at *http://115.167.2.211/canpro-idss/* and the top page of the CanPro-IDSS system will be displayed on your computer screen (Fig. 4.2a).

II. An "Example" link displays the format of query sequences. Datasets of breast cancer and Non-breast cancer related protein sequences are attached under the "Dataset" link. Enter at least two query proteins in fasta format into text box below the statement "*Enter sequences in fasta format (at least 2 sequences)*" given in the displayed page. After entering query proteins as shown in (Fig. 4.2b), click on the "*Submit*" button to see the predicted output.

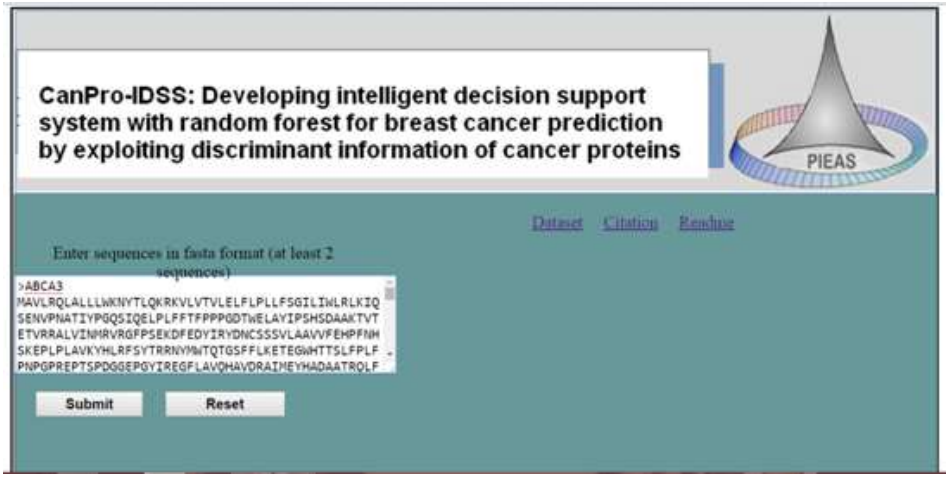
III. When the *Submit* button is clicked, the CanPro-IDSS system initiates preprocessing before executing the results. It first check for illegal characters. After checking the constraints, the CanPro-IDSS system determines the query as a BC/NBC proteins, as illustrated in Fig. 4.2c.

4.2 Experiment Framework

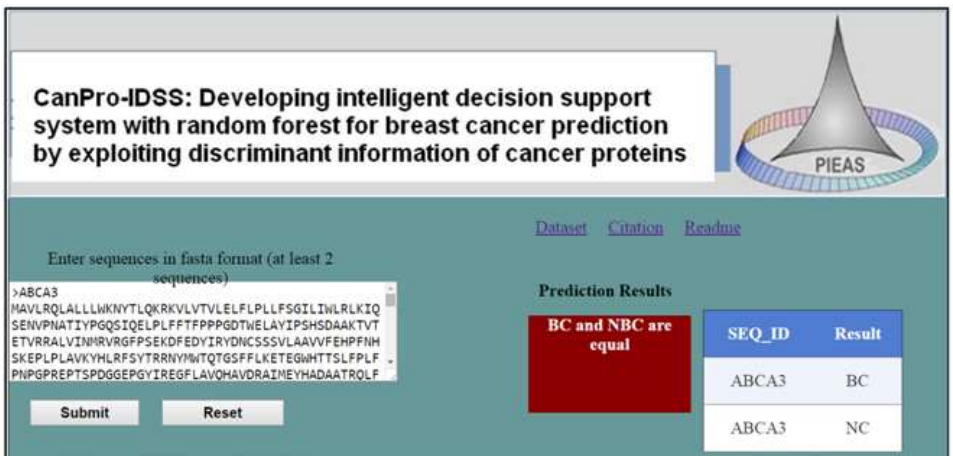
In given dataset (Table 2.1) non-cancer patients outnumber the cancer patients. When algorithms are implemented to the imbalanced dataset, they are overwhelmed by examples in the majority class and over looked the examples of the minority class. Consequently, learning algorithms are generating high performance for the majority class and weak performance for the minority class. In this scenario, MTD and CSC approaches are utilized. The implementations of MTD with RF, AdaBoostM1, Bagging, and GentleBoost are shown in Fig.4.1a. However, implementations of CSC with GentleBoost, AdaBoostM1, and Bagging are shown in Fig. 4.1b. Ten folds cross-validation is employed to assess the performances. The average of Sn, Sp, Acc, G_{mean}, and AUC are calculated for 10 individual folds.



(a)



(b)



(c)

Figure 4.2 Screenshot (a) Display demonstrates the Main page of CanPro-IDSS cancer prediction system (b) Display illustrating the input query of protein sequences, and (c) Display showing the predicted type of proteins as a BC or NBC.

MTD is implemented in feature space to enhance the examples of minority class by generating diffuse examples. Class imbalance ratios are 4.53 and 7.66 for datasets of C/NC and B/NBC, respectively (Table 2.1). So, the amount of over-sampling is adjusted according to class imbalance ratio. This ratio made one and thereby balanced the numbers of examples in majority and minority. However, CSL is implemented to the imbalanced training dataset. During training, CSC is reweighted each training example according to given costs. The misclassification cost is assigned to train model as proposed in [68]; explicitly, the cost of misclassifying a true class (breast cancer) as false (non-breast cancer) comes equal to the imbalance ratio. Therefore, we assigned $C_t(1,0)=4.53$ and $C_t(1,0)=7.66$ for datasets of C/NC and B/NBC, respectively. The cost of misclassifying a non-cancer as cancer is given to $C_t(0,1)=1$. The cost of true classification is set to 0, i.e., $C_t(1,1)= C_t(0,0)=0$.

RF, GentleBoost, AdaBoostM1, and Bagging are implemented using imbalanced dataset. Boosting utilizes the whole dataset to train base predictors in serial. However, Bagging technique sample training data to generate diverse predictors and combined their decisions to build the final decision. Bagging ensemble is implemented using discriminant analysis, whereas, AdaBoostM1 and GentleBoost are implemented using DT. For the selection of a suitable number of classifiers of ensemble system, different curves of GentleBoost, AdaBoostM1,and Bagging are constructed using various feature spaces. Fig. 4.3 shows the ensemble error as a function of the number of trees in the ensemble using PseAAC-S feature space for C/NC and B/NBC datasets. This figure highlights that ensemble error asymptotically decreases to a certain height as the size of ensemble increases.

On the other hand, RF models are developed by utilizing random sub-sampling of the input dataset, and then employing the ensemble strategies to achieve accurate predictions. Generalization of RF comes from Bagging scheme which makes better generalization by decreasing variance and use of Boosting scheme helped in decreasing bias. RF ensemble is trained in different feature spaces by the varying number of trees. In RF, two training parameters: (i) the number of trees and (ii) the number of randomly selected variables i.e., *mtry* to generate optimal model are used.

Fig. 4.4 indicates the performance accuracy of RF ensemble with the increasing number of trees (*ntree*) using different feature spaces. The accuracy of RF

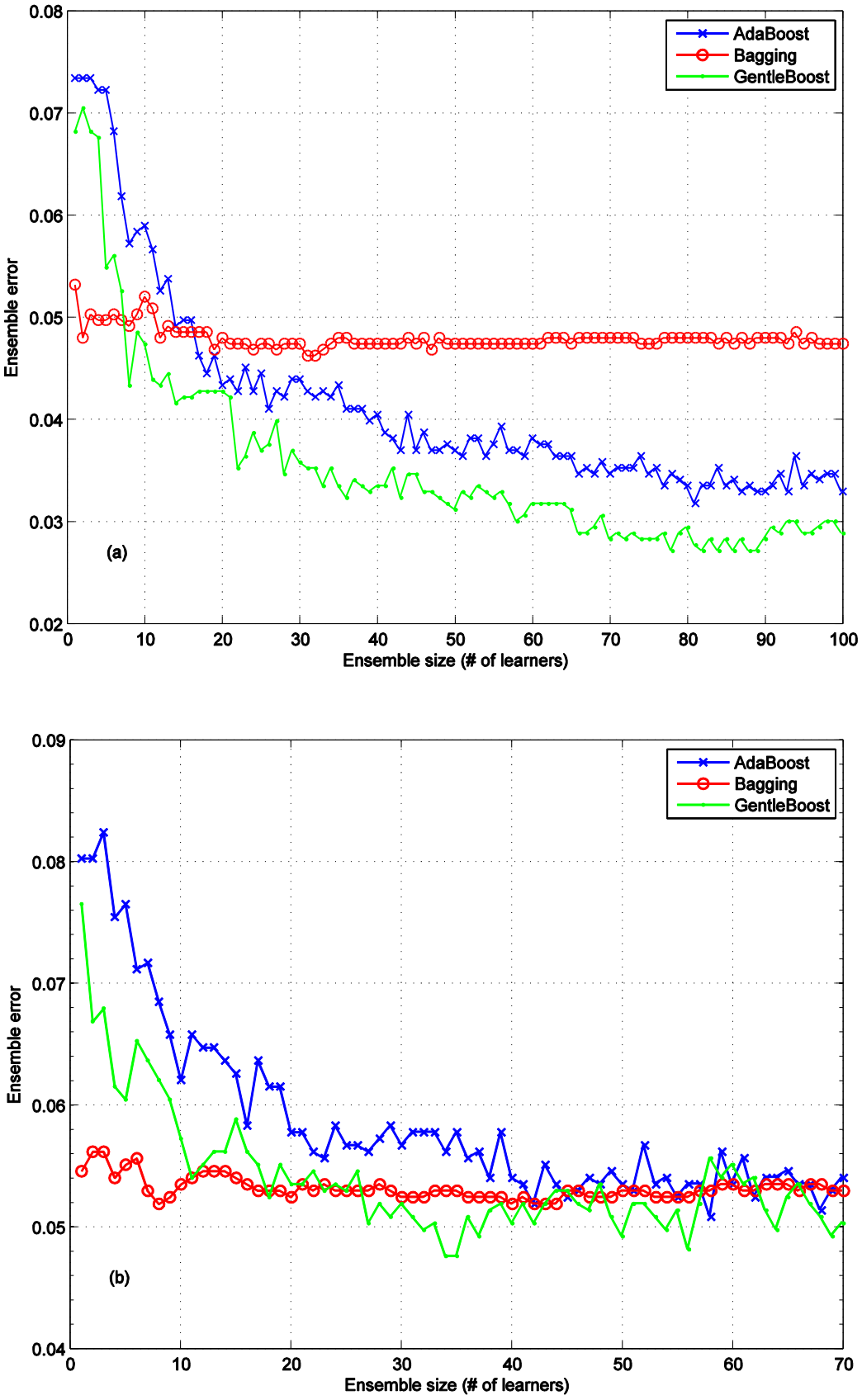


Figure 4.3 Ensemble error as a function of the number of learners in the ensembles of GentleBoost, AdaBoostM1, and Bagging using PseAAC-S feature space for (a) C/NC and (b) B/NBC datasets.

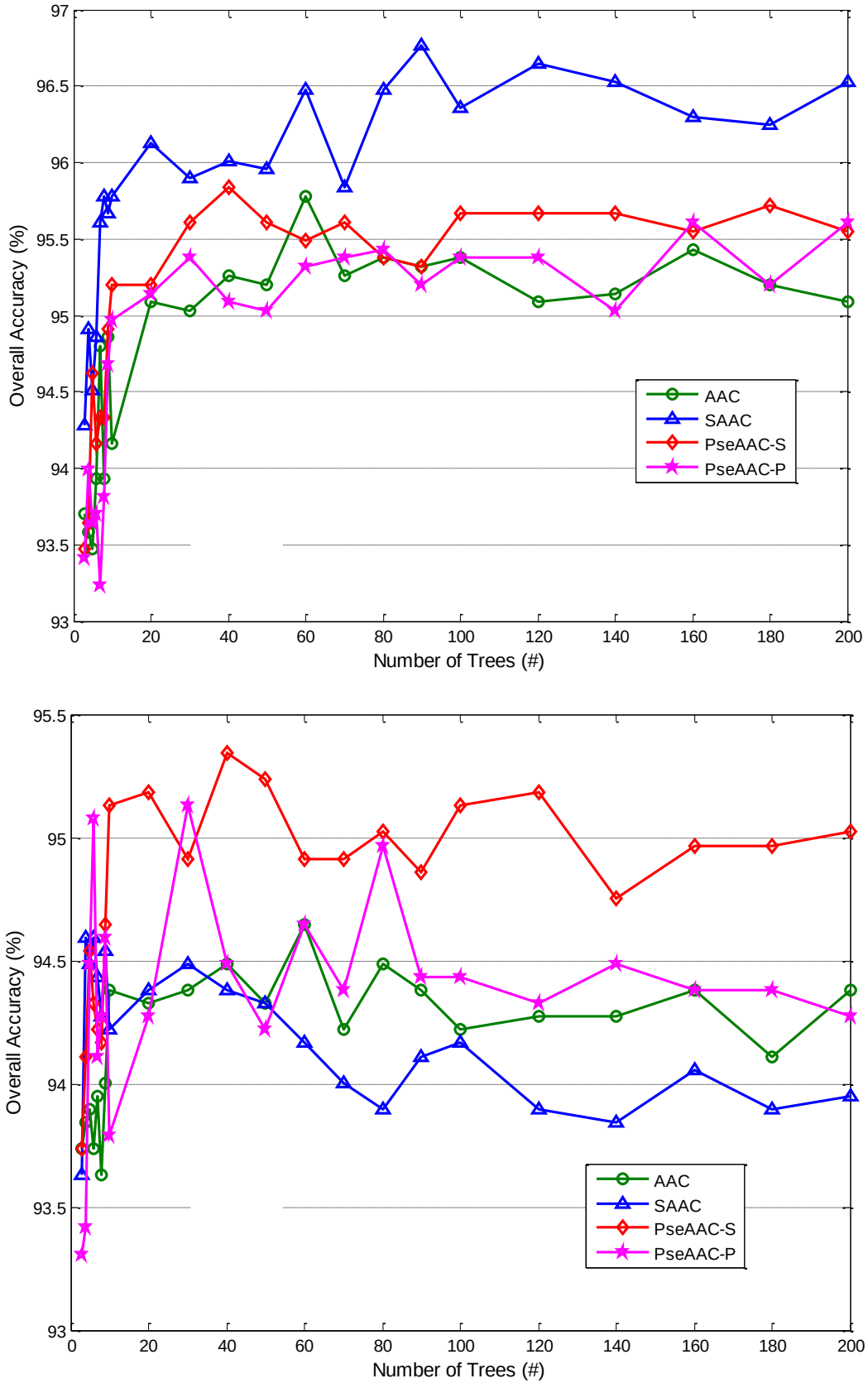


Figure 4.4 Prediction accuracies of RF vs. number of trees using different feature spaces of AAC, SAAC, PseAAC-S, and PseAAC-P for (a) C/NC and (b) B/NBC datasets.

ensemble by varying *ntree* from 1 to 200 with equal interval of 20 trees is computed. Fig. 4.4a shows that initially the performance of RF ensemble is considerably improved with the increase of the number of trees. Beyond a certain limit (20-80), there is no appreciable change in overall accuracy. It is observed that SAAC feature space has provided the best overall accuracy of 96.76% for 90 trees, followed by PseAAC-S feature space with 95.84% of accuracy for 40 trees. However, AAC and PseAAC-P feature spaces have given the overall accuracies of 95.78% and 95.61% for 60 and 160 trees, respectively. Fig. 4.4b depicts overall accuracies against the number of trees in various feature spaces for B/NBC dataset. It is observed that PseAAC-S feature space has provided the best overall accuracy of 95.34% for 40 trees. While, PseAAC-P feature space has yielded an accuracy of 95.13% for 30 trees.

4.3 Results and Discussion

Several experiments are performed to explore the effectiveness of the proposed systems. Three sets of ensemble models are developed using: (i) without MTD and CSC, (ii) CSL technique, and (iii) MTD technique for different feature spaces of AAC, SAAC, PseAAC-S and PseAAC-S. The performance of these ensemble models is discussed as:

4.3.1 Performance of Models Without MTD and CSL

Table 4.2 highlights the performance comparison of RF, AdaBoostM1, Bagging, and GentleBoost ensembles approaches without employing MTD and CSL techniques (i.e., original datasets). It is observed that, for C/NC dataset, RF model has given the highest values of Acc 94.32% and 93.47% for AAC and PseAAC-P feature spaces, respectively. However, for other feature spaces, GentleBoost ensemble has provided the values of Acc 93.84% and 94.22%, for SAAC and PseAAC-S feature spaces respectively. Overall, all the prediction models performed better for PseAAC-S feature space and yielded average Acc near to 93.66%. Form Table 4.2, it is found that RF model has better decision than AdaBoostM1, Bagging, and GentleBoost for B/NBC dataset. However, it is observed that for both datasets, the value of Sp is lower than Sn value. This is due to the imbalanced nature of the input data, as more the imbalanced data the lower the value of Sp for B/NBC dataset. Table 4.3 shows the performance, in terms of AUC, of conventional classification models of RF, AdaBoostM1, Bagging, and GentleBoost models using imbalanced datasets.

Table 4.2 Performance comparison of the models without MTD and CSL techniques.

Model/Feature space		C/NC dataset					B/NBC dataset				
		Acc	Sp	Sn	G_{mean}	F_{score}	Acc	Sp	Sn	G_{mean}	F_{score}
AAC	RF	94.32	87.43	95.84	91.54	96.51	94.32	87.43	95.84	91.54	96.51
	AdaBoostM1	87.78	53.93	95.26	71.67	92.74	88.26	9.84	98.50	31.13	93.69
	Bagging	89.30	60.73	95.61	76.20	93.60	89.87	40.98	96.25	62.81	94.38
	GentleBoost	88.73	59.69	95.14	75.36	93.26	88.73	9.84	99.04	31.21	93.96
SAAC	RF	93.37	86.39	94.91	90.55	95.91	93.37	86.39	94.91	90.55	95.91
	AdaBoostM1	92.80	75.39	96.65	85.36	95.65	88.54	15.57	98.07	39.08	93.80
	Bagging	90.81	69.63	95.49	81.54	94.45	91.38	54.92	96.15	72.66	95.18
	GentleBoost	93.84	79.06	97.11	87.62	96.28	90.06	32.79	97.54	56.55	94.55
PseAAC-S	RF	93.28	84.82	95.14	89.83	95.86	93.28	84.82	95.14	89.83	95.86
	AdaBoostM1	94.03	81.68	96.76	88.90	96.37	91.57	59.84	95.72	75.68	95.26
	Bagging	93.09	81.68	95.61	88.37	95.77	90.63	59.84	94.65	75.25	94.70
	GentleBoost	94.22	80.63	97.23	88.54	96.50	92.14	59.84	96.36	75.93	95.59
PseAAC-P	RF	93.47	87.96	94.68	91.26	95.96	93.47	87.96	94.68	91.26	95.96
	AdaBoostM1	91.29	68.06	96.42	81.01	94.77	89.20	13.11	99.14	36.06	94.20
	Bagging	90.81	65.97	96.30	79.70	94.50	90.63	46.72	96.36	67.10	94.79
	GentleBoost	90.91	68.06	95.95	80.81	94.53	89.02	33.61	96.25	56.87	93.94

Table 4.3 Performance comparison, in terms of AUC, of RF, AdaBoostM1, Bagging, and GentleBoost ensemble approaches for imbalanced datasets.

Dataset/Model	C/NC dataset				B/NBC dataset			
	AAC	SAAC	PseAAC-S	PseAAC-P	AAC	SAAC	PseAAC-S	PseAAC-P
RF	97.53	97.32	97.44	96.80	95.46	96.83	95.80	92.64
AdaBoostM1	87.73	95.16	97.41	93.07	76.84	86.52	93.08	88.81
Bagging	91.54	92.73	95.78	93.96	89.23	90.36	92.87	91.52
GentleBoost	89.94	96.28	97.30	93.4	76.52	90.89	92.52	89.83

4.3.2 Performance of Can-CSCGnB System

The performance comparison of CSC-GentleBoost with CSC-AdaBoostM1, and CSC-Bagging is provided in Table 4.4. It is observed that, for C/NC dataset, CSC-GentleBoost has given the best value of Acc 94.41% for PseAAC-S feature space. It has also provided Acc values 86.27%, 90.53%, and 90.34% for AAC, SAAC, and PseAAC-P feature spaces, respectively. For B/NBC dataset, It is observed that CSC-GentleBoost has provided best overall Acc 90.15% in PseAAC-S feature space. It has also yielded Acc values of 85.51%, 89.87%, and 88.64 % for AAC, SAAC, and PseAAC-P feature spaces, respectively. It is inferred that the Can-CSCGnB system performed better for both datasets. However, it observed that employing CSC approach, the values of Sp and G_{mean} are enhanced and the values of Acc and Sn are

reduced for both datasets. For C/NC dataset, on average, CSC-GentleBoost has increased the values of Sp (12.04%), and G_{mean} (4.68%), whereas the values of Acc (-1.54%) and Sn (-4.54%) is reduced. This change is more profound for B/NBC dataset. On average, Can-CSCGnB system has the raised values of Sp (38.32%), and G_{mean} (25.75%), whereas the values of Acc (-1.45%) and Sn (-6.64%) is lowered. In medical application, an improved ROC curve is always required. This is achieved with the higher values of Sn and Sp. This improved performance highlights the effectiveness of CSC approach for the prediction of cancer.

Table 4.4 Prediction performance of the CSC based models.

Model/Feature space		C/NC dataset				B/NBC dataset			
		Acc	Sp	Sn	Gmean	Acc	Sp	Sn	Gmean
AAC	CSC-AdaBoostM1	83.61	79.58	84.51	82.01	81.82	70.49	83.30	76.63
	CSC-Bagging	83.62	85.34	83.24	84.28	80.30	82.79	79.98	81.37
	CSC-GentleBoost	86.27	82.20	87.17	84.65	85.51	68.03	87.79	77.28
SAAC	CSC-AdaBoostM1	87.97	89.01	87.75	88.37	87.78	81.15	88.65	84.82
	CSC-Bagging	86.84	88.48	86.47	87.47	83.24	84.43	83.08	83.75
	CSC-GentleBoost	90.53	86.39	91.45	88.88	89.87	83.61	90.69	87.07
PseAAC-S	CSC-AdaBoostM1	93.56	88.48	94.68	91.53	89.02	79.51	90.26	84.71
	CSC-Bagging	88.35	95.29	86.82	90.96	84.47	93.44	83.30	88.22
	CSC-GentleBoost	94.41	85.34	96.42	90.71	90.15	72.13	92.51	81.69
PseAAC-P	CSC-AdaBoostM1	86.93	81.68	88.09	84.82	87.31	67.21	89.94	77.75
	CSC-Bagging	86.08	88.48	85.55	87.00	83.62	86.07	83.30	84.67
	CSC-GentleBoost	90.34	81.68	92.25	86.80	88.64	65.57	91.65	77.52

ROC curves of CSC based models using AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces are depicted in Fig. 4.5. It is observed that for C/NC dataset (Fig. 4.5A (a-d)), all models have provided higher AUC (above 97%) in PseAAC-S feature space. However, for B/NBC dataset (Fig. 4.5B (a-d)), CSC-GentleBoost has shown the best performance (94.28%) in SAAC feature space, followed by CSC-Bagging (94.27%) in PseAAC-S and CSC-AdaBoostM1 (93.34%) in SAAC space. On the other hand, in terms of AUC measure, an average enhancement of 2.14, 0.60, and 1.78 % is observed for CSC-AdaBoostM1, CSC-Bagging, and CSC-GentleBoost ensemble models, respectively. CSL takes care of different misclassification costs of false negatives and false positives, the enhancement in performance demonstrated its effectiveness to handle imbalanced data. Thus, it is an efficient technique, which reduce the total misclassification cost of models.

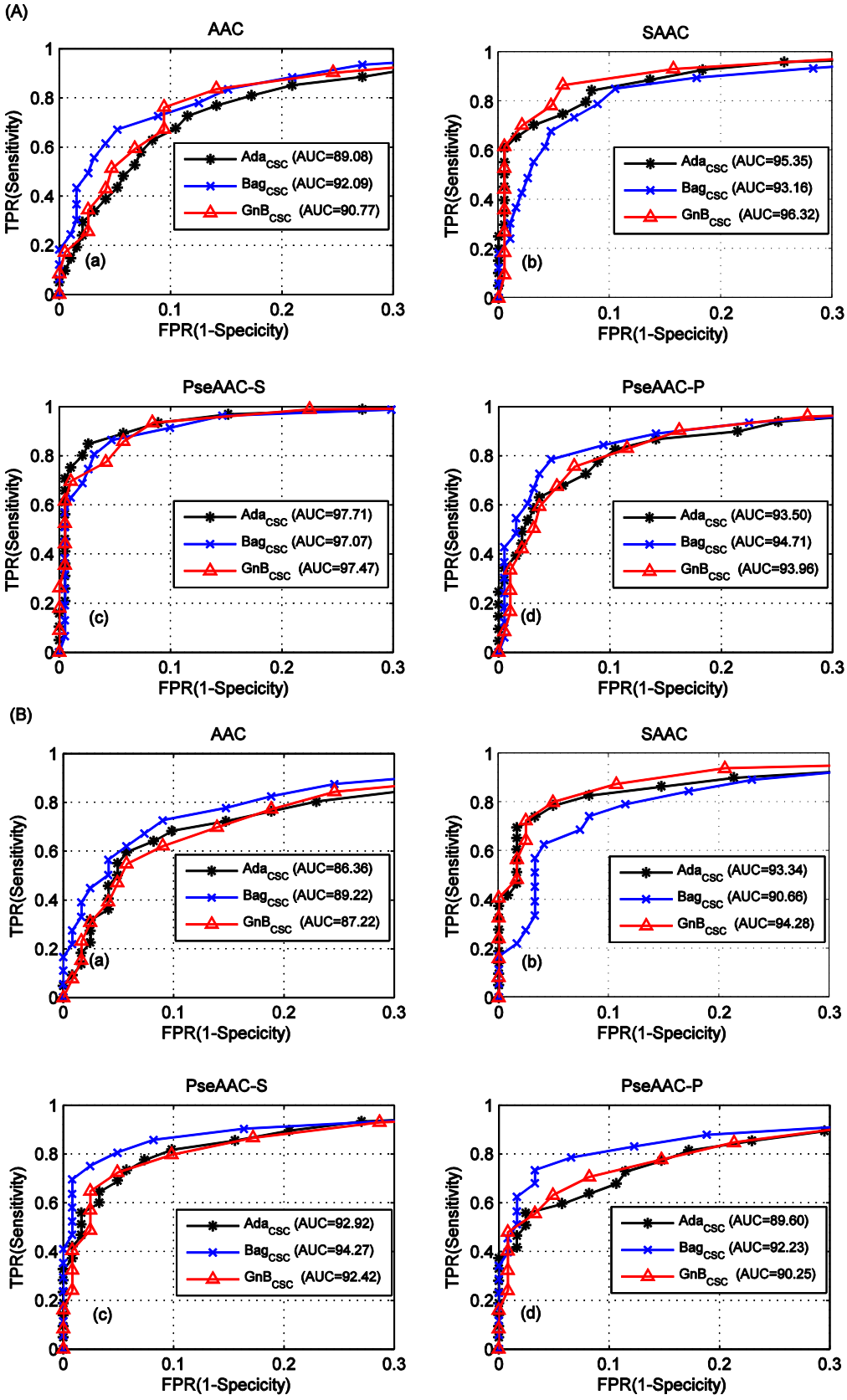


Figure 4.5 ROC curves of CSC-AdaBoostM1, CSC-Bagging, and CSC-GentleBoost (GnB) approaches for (A) C/NC and (B) B/BNC datasets using: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P spaces. ***Note that for better visualization of region of interest, partial ROC curves are plotted.***

4.3.3 Performance of the Proposed CanPro-IDSS System

Table 4.5 demonstrates the overall comparison of MTD-RF, MTD-AdaBoostM1, MTD-Bagging, and MTD-GentleBoost models using the MTD technique. It is observed that, for C/NC dataset, MTD-RF has achieved the best Acc of 97.66% for PseAAC-S space. MTD-RF has also provided better Acc of 96.85, 96.76, and 96.91 % for AAC, SAAC, and PseAAC-P spaces, respectively. For B/NBC dataset, it is observed that again MTD-RF has attained the best overall accuracy of 97.10 % in PseAAC-P feature space. It has also yielded Acc of 96.52, 96.43, and 96.78 % for AAC, SAAC, and PseAAC-S spaces, respectively. It is noted that the proposed MTD-RF model has given improved values for all measures. For C/NC dataset, on average, MTD-RF has increased the values of Acc (3.44%), Sp (10.84%), Sn 1.45%), and G_{mean} (6.22%). For B/BNC dataset, on average, MTD-RF has enhanced the values of Acc (3.09%), Sp (10.52%), Sn (1.11%), and G_{mean} (5.95%). It is inferred that MTD-RF performed better for both datasets. Here, it is noted that MTD-RF model based the complete cancer system is denoted by CanPro-IDSS.

On average, all classification models have reported their increased performance. However, MTD-AdaboostM1 and MTD-GentleBoost models have reduced the value of Sn for some spaces. The enhancement of models, specifically RF, highlights the effectiveness of MTD technique for the prediction of cancerous protein molecule.

In medical decision, an ameliorated ROC curve is beneficial for the selection of operating point of a classifier. Fig. 4.6 shows ROC curves of MTD based models for AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces. Fig. 4.6A(a-d) and Fig. 4.6B(a-d) have highlighted an improved ROC curves for C/NC and B/NBC datasets, respectively. Form Fig. 4.6, it is observed that MTD-RF has better decision in PseAAC-S space than other models for cancerous protein molecules.

For C/NC dataset, comparison of Table 4.3 and Fig. 4.6 show that MTD-RF model has achieved improvement of 1.71%, 2.35%, 2.13%, and 2.41% in AAC, SAAC, PseAAC-S, and PseAAC-P spaces, respectively. Whereas, for B/NBC dataset, Tables 4.3 and Fig. 4.6 indicate that MTD-RF has provided an enhancement of 3.61%, 2.14%, 3.39% and 6.55% in AAC, SAAC, PseAAC-S, and PseAAC-P spaces, respectively. As a result, when MTD approach is applied, an average enhancement of

2.89%, 8.04%, 6.29%, and 7.37% for MTD-RF, MTD-AdaBoostM1, MTD-Bagging, and MTD-GentleBoost models, respectively is obtained. Considerable enhancement is obtained in ROC curves of the balanced dataset. Therefore, MTD technique is very successful for imbalance data. The comparison summarized that CanPro-IDSS is the best approach for the prediction of cancer protein molecules. The proposed approach outperformed the conventional ensemble approaches and CSC based models.

Table 4.5 Performance of the proposed MTD based models.

Model/Feature space		C/NC dataset				B/NBC dataset			
		Acc	Sp	Sn	G_{mean}	Acc	Sp	Sn	G_{mean}
AAC	MTD-RF	96.85	96.75	96.77	96.78	96.52	97.33	95.60	96.45
	MTD-AdaBoostM1	90.69	90.31	91.18	90.69	90.60	93.31	88.00	90.62
	MTD-Bagging	93.38	90.31	96.54	93.42	93.96	92.53	95.39	93.89
	MTD-GentleBoost	92.81	91.19	94.48	92.81	92.85	93.01	92.73	92.90
SAAC	MTD-RF	96.76	97.31	96.18	96.75	96.43	96.86	96.04	96.52
	MTD-AdaBoostM1	93.37	90.32	96.54	93.42	90.53	90.61	90.44	90.52
	MTD-Bagging	94.49	91.30	97.72	94.50	94.39	90.03	98.78	94.31
	MTD-GentleBoost	95.30	93.47	97.10	95.29	92.89	91.48	94.32	92.91
PseAAC-S	MTD-RF	97.66	98.17	97.13	97.63	96.78	97.63	96.03	96.79
	MTD-AdaBoostM1	96.43	95.34	97.52	96.43	91.90	93.22	90.73	91.89
	MTD-Bagging	95.25	92.79	97.78	95.30	94.52	91.00	98.10	94.52
	MTD-GentleBoost	96.12	95.50	96.71	96.11	93.80	91.69	95.69	93.69
PseAAC-P	MTD-RF	96.91	97.73	96.29	96.89	97.10	96.86	97.34	97.20
	MTD-AdaBoostM1	94.41	93.22	95.64	94.41	93.51	91.83	95.23	93.51
	MTD-Bagging	94.19	90.47	97.89	94.22	94.69	91.23	98.30	94.70
	MTD-GentleBoost	94.61	93.10	96.13	94.49	93.31	91.10	95.39	93.22

The analysis of variance (ANOVA) statistical test was conducted to demonstrate which model and feature space are performing significantly different from others. Table 4.6 reveals the ANOVA results in terms of mean-Acc and AUC-ROC for C/NC and B/NBC datasets, respectively. This analysis shows significant difference among MTD-RF models, since the yielded P-values 0.0011 and 0.0001 are lower than $\alpha = 0.05$ for the C/NC and B/NBC datasets, respectively. In case of B/NBC dataset, the P-value is zero to three decimal places. This signifies that the performance, in terms of AUC-ROC, varies from one MTD-RF model to another. Additionally, we implemented multiple comparison procedures to investigate the significant difference between each pair of models and feature spaces. The graph

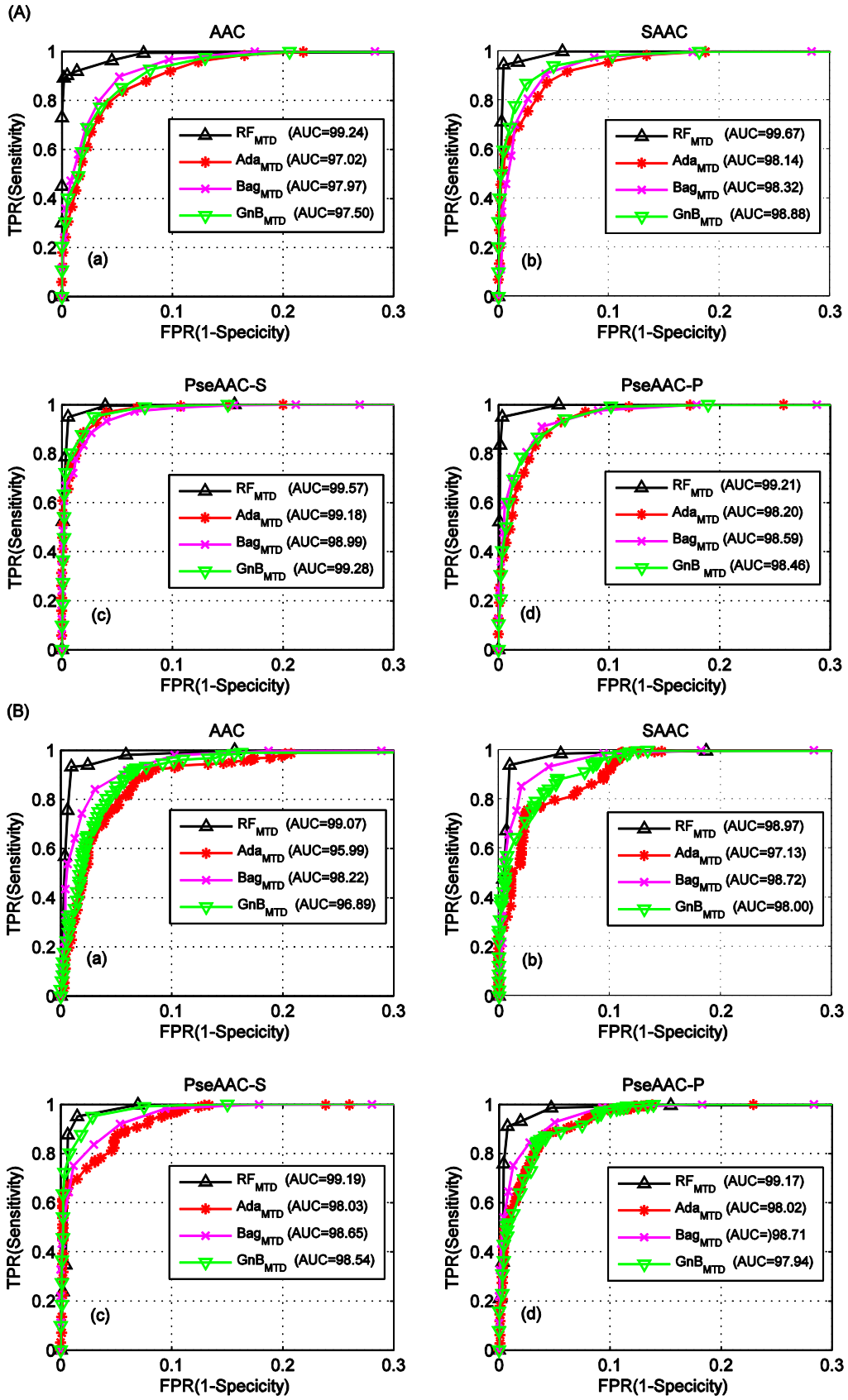


Figure 4.6 ROC curves of MTD-RF, MTD-AdaBoostM1, MTD-Bagging, and MTD-GentleBoost models for (A) C/NC and (B) B/BNC datasets using: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P spaces.

Table 4.6 ANOVA test for mean-Acc and AUC using C/NC and B/NBC datasets.

Dataset	Source	Sum of squares	DF	Mean square	F	*P*-value
C/NC	Models	2.6609	3	0.8869	7.17	0.0011
	Rows (Acc)	1.2991	9	0.1443	1.17	0.3539
	Error	3.3389	27	0.1236		
	Total	7.2989	39			
B/NBC	Models	2.5083	3	0.8361	48.41	0.0001
	Rows (AUC)	0.1245	9	0.0138	00.80	0.6188
	Error	0.4663	27	0.0172		
	Total	3.0990	39			

(Fig.4.7a) of the multiple comparison tests highlights that PseAAC-S feature space, in terms of Acc, is only significantly different from AAC, SAAC, and PseAAC-P feature spaces for the C/NC dataset. Fig.4.7b shows that PseAAC-S feature space, in terms of AUC-ROC, is significantly different from AAC and SAAC feature spaces for B/NBC dataset. The overall performance comparison of the proposed approach with previous studies is given in the next section.

4.3.4 Overall Performance Comparison

Fig. 4.8 highlights relative improvement in performance measures of the proposed CanPro-IDSS over CSC based models and conventional approaches. For C/NC dataset, Fig. 4.8A(a-d) accentuates that our approach has gained the highest RIA in Sn measure of 36.01 % with CSC-Bagging, and 51.91% with CSC-AdaBoostM1. In terms of Sp, the highest RIA over the conventional ensemble of 50.15% with RF, 172.26 % with AdaBoostM1, 167.4% with Bagging, and 150.52% with GentleBoost, and over CSC based models of 61.5% with CSC-AdaBoostM1, 36.83% with CSC-Bagging, and 65.02% CSC-GentleBoost is observed. For Acc and AUC, the highest RIA of 50.36% with CSC-Bagging and 26.67% with AdaBoostM1 is observed, respectively. Similarly, for B/NBC dataset, Fig. 4.8B(a-d), it is observed that the proposed approach has attained the highest RIA over conventional models of AdaBoostM1, Bagging, and GentleBoost in Sn, Sp, Acc, and AUC than CSC based models. The Sp and Sn of classification is considerable improved by employing MTD function. The higher values of Sn and Sp are always required for better decision.

The conventional prediction systems AdaBoostM1, Bagging, and GentleBoost

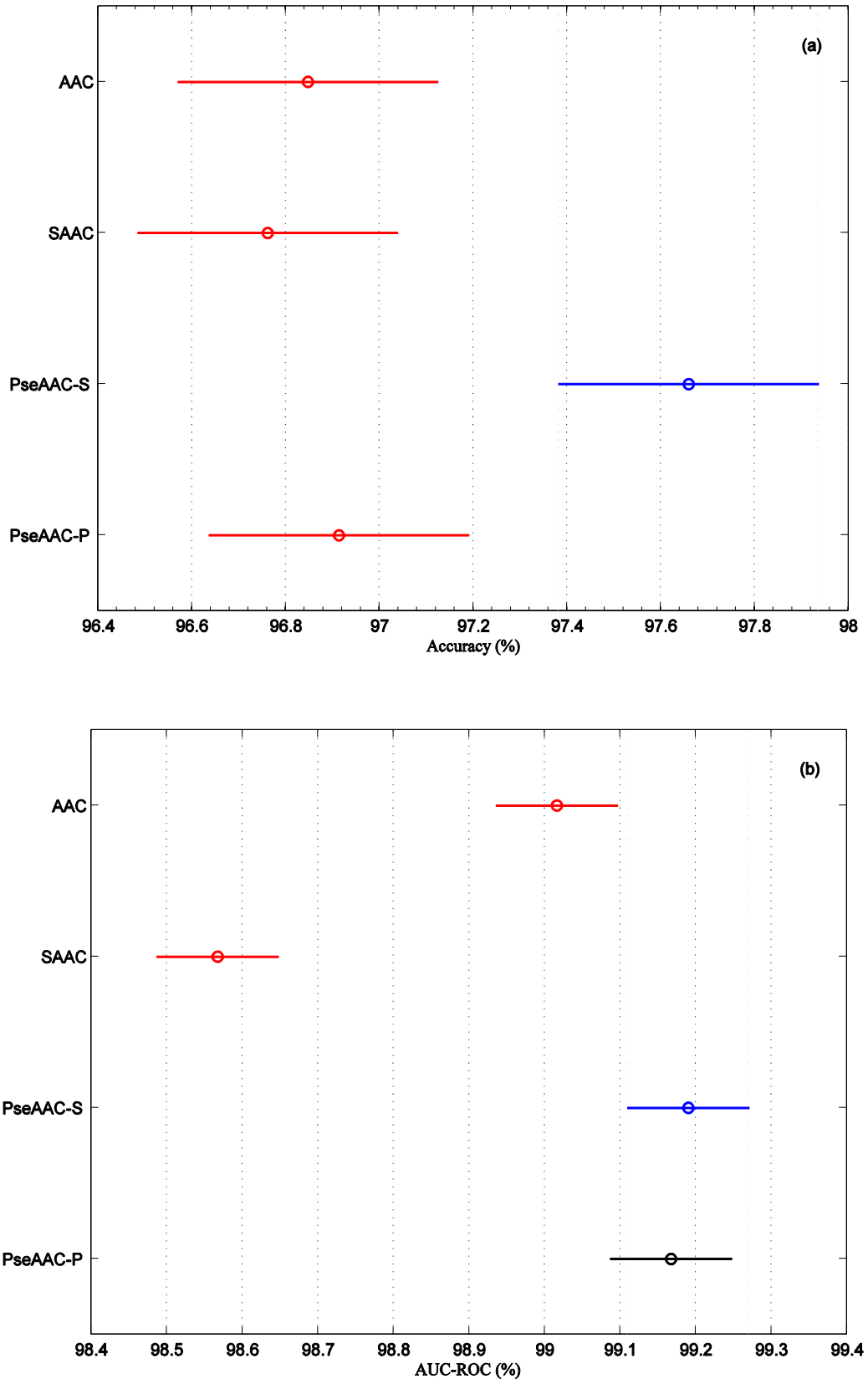


Figure 4.7 Multiple comparison tests of mean values of (a) accuracy for C/NC dataset and (b) area under the curve of ROC for B/NBC dataset in different models and feature spaces.

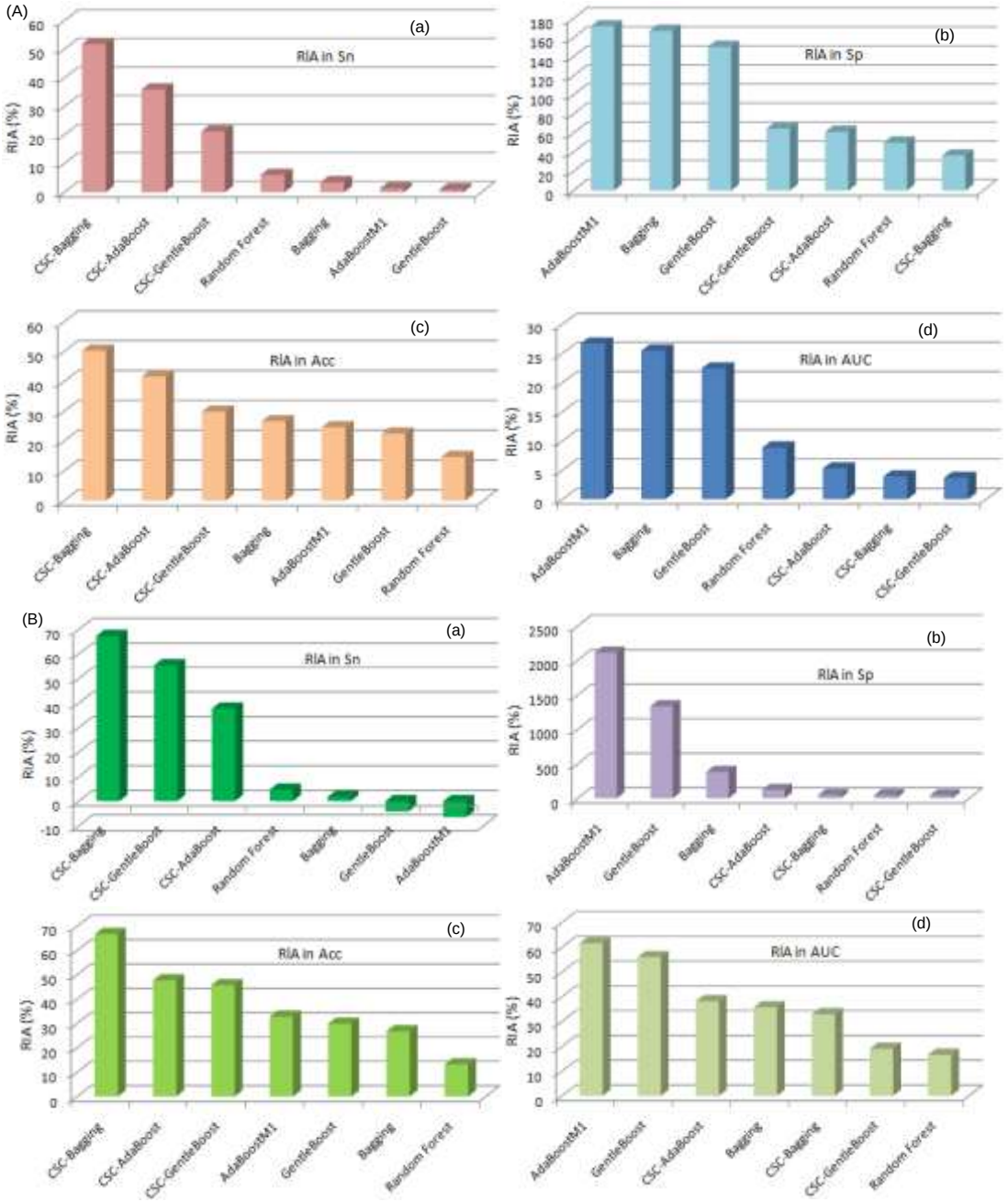


Figure 4.8 RIA of the proposed MTD-RF based approach in performance measures of Sn, Sp, Acc, and AUC for (A) C/NC and (B) B/BNC datasets.

have shown reduced performance, because they could not generate better decision space for the prediction of cancer. On the other hand, the proposed random ensemble system is excellent over other approaches because: (i) it incorporates MTD as preprocessor for data balancing, and (ii) RF employs random data sub-sampling and ensemble strategies to generate better decision space.

In Table 4.6, a performance comparison of the proposed MTD and CSC based approaches with previous approaches is carried out. Wang et al. applied imbalanced

data for the prediction of the survivability prognosis of breast cancer [68]. They attained maximum AUC of 84.2% and 82.0% for logistic regression (LR) algorithm (C_rLR) and DT algorithm (C_pDT), respectively. Delen et al. used surveillance and epidemiology results for the prediction of breast cancer [70]. They employed three classification models of DT, LR, and ANN. They reported the highest AUC of 84.9% and 76.9% for LR and DT, respectively. Zhang et al. utilized gene expression profiles for the prediction of breast cancer by employing LR, SVM, AdaBoost, LogitBoost and RF [69]. They achieved maximum AUC of 88.6% and 89.9% for SVM and RF models, respectively. However, we obtained breast cancer prediction of MTD-RF 99.2% (Table 4.7). In another study, Khalilia et al. developed prediction models from highly imbalanced data using SVM, Bagging, Boosting and RF [71]. They demonstrated that, in terms of AUC, RF model (91.2%) outperformed SVM (90.6%), Bagging (90.5%), and Boosting (88.9%).

Table 4.7 Prediction comparison in terms of AUC of the proposed approach CanPro-IDSS with previous approaches for breast cancer.

Approach	AUC (%)	Acc	Sp	Sn	Dataset
C_rLR	84.2	75.1	75.0	76.2	Surveillance, Epidemiology, and End Results (SEER) data [68]
C_pDT	82.0	75.8	75.8	75.6	-DO-
DT	76.9	90.3	98.4	27.9	SEER data [70]
LR	83.2	89.7	98.8	22.6	SEER data [104]
SVM	87.4	NA	NA	NA	van de Vijver [105] and Wang [106] datasets [69]
RF	93.2	NA	NA	NA	-Do-
RF	91.2	NA	NA	NA	Nationwide Inpatient Sample (NIS) database [71]
SVM	90.6	NA	NA	NA	-Do-
Bagging	90.5	NA	NA	NA	-Do-
Boosting	88.9	NA	NA	NA	-Do-
CSC-AdaBoostM1	93.34	89.02	81.2	90.3	Present study
CSC-Bagging	94.27	84.47	93.4	83.3	-Do-
CSC-GentleBoost	94.28	90.15	83.6	92.5	-Do-
MDT-AdaBoostM1	98.0	93.5	93.3	95.2	-Do-
MDT-Bagging	98.7	94.5	92.5	98.8	-Do-
MDT-GentleBoost	98.5	93.8	93.0	95.7	-Do-
Proposed approach	99.2	97.1	97.6	97.3	-Do-

On the other hand, the proposed approach using Hd and Hb properties of amino acids has given the best AUC of 99.7% for cancer prediction (Fig. 4.6A(b)) and 99.2% for breast cancer (Fig. 4.6B(d)). Other approaches of MTD-AdaBoostM1, MTD-Bagging, and MTD-GentleBoost have provided AUC in the range of 98% for breast cancer.

4.3.5 Computational Time Based Comparison

In order to explore the efficient model/predictor, we compared the computation time caused by MTD between KNN, SVM, NB, (Chapter 3) and RF predictors. The experimental work is carried out using Intel(R) dual-core, 2.93GHz, 2GB RAM with Windows 7 Operating System. The MTD is used to increase minority class instances, thereby, balancing three types of cancer datasets. Consequently, predictors have achieved better level of prediction accuracy at the expense of computing cost. Fig. 4.9 demonstrates the computational time comparisons of KNN, SVM, NB, and RF predictors for C/NC, B/NBC, and CC/NCC datasets using different feature spaces. This figure highlights the difference in computational time of different predictors after and before MTD technique. It is noticed that all predictors have consumed the least computation time in AAC feature space, while the most computation time is taken by PseAAC-S feature space.

This figure reflects that MTD-RF predictor is relatively slower than MTD-KNN and other predictors. Because, RF is a tree-based ensemble classifier that computes the prediction by utilizing two stages of random feature subspace and out-of-bag estimates. However, MTD-RF has shown comparable prediction performance with MTD-SVM (see Table 3.10). On the other hand, KNN approach uses majority votes of different K neighbors to classify input sequence. MTD-KNN model has achieved moderately better prediction accuracy as compared to MTD-NB, but at the expense of higher computing cost. MTD-KNN is faster than MTD-RF, its prediction performance is not better than MTD-RF (see Table 3.10 and Table 4.5).

Fig. 4.9 shows that, in general, NB predictor is faster than other predictors, but it has less overall prediction performance (see Table 3.10). Overall, the computing time of MTD-SVM predictor is comparable with MTD-NB for B/NBC and CC/NCC datasets in all feature spaces, expect PseAAC-S. With the highest accuracy performance, MTD-SVM is emerged as efficient predictor. It is observed that, with

balanced data, on the average SVM is 3.95 and 12.13 times more efficient than KNN and RF for C/NC dataset, respectively. However, SVM is 2.72 and 5.82 times more efficient than KNN and RF for B/NBC dataset, respectively. Therefore, MTD-SVM predictor is more appealing in terms of prediction performance and computational time. The generalization strength of SVM is higher and to address non-linear prediction, that enables to perform well on novel data. Though, the synthetic data generation is slightly expensive. But SVM predictor is very effective in learning from balanced dataset.

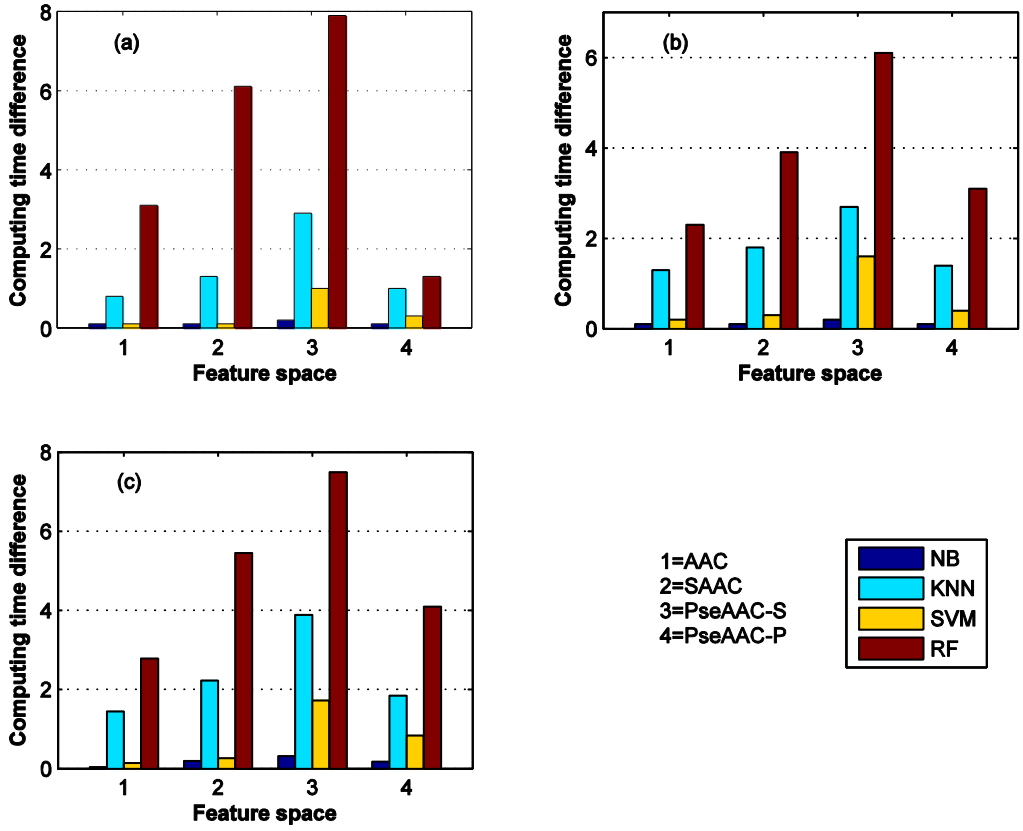


Figure 4.9 Computational time comparisons of KNN, SVM, NB, and RF for (a) C/NC, (b) B/NBC, and (c) CC/NCC datasets using feature spaces of different dimensions.

In this chapter, homogeneous ensemble systems are developed by combining MTD or CSC with different ensemble approaches of RF or GentleBoost. The results highlighted that the proposed random ensemble system, CanPro-IDSS, using balanced data outperformed over the CSL based system and previous approaches. In the next chapter, we shall combine the best data balancing technique with prediction models. For this purpose, SVM, KNN, NB, DT, and PNN learning approaches are selected.

Chapter 5: Improving Prediction by Ensemble Voting Strategy

The main objective of this chapter is to discuss a novel heterogeneous decision making ensemble system "IDMS-HBC" for the prediction of human breast cancer. The proposed system is achieved by the integration of diverse predictions of base predictors through non-trainable majority voting strategy. In this work, the learning mechanisms of SVM, KNN, NB, DT, and PNN are selected as base predictors. Every learning model has different inductive bias and learning hypotheses such as instance-based, trees, probability, and statistics. Thus, each model provides potentially independent and diverse predictions. The overview of the proposed system is provided as follows:

5.1 The Proposed IDMS-HBC System

The basic block diagram of the proposed IDMS-HBC system is shown in Fig. 5.1. The proposed system has three main modules. In the first stage, useful information is extracted from protein primary sequences using physicochemical properties of amino acids of Hd and Hb. Databases of feature spaces are scaled between [0, 1] in order that features with large numeric values could not dominate with small numeric values. In data balancing module, the MTD technique is utilized to oversampling the minority class, so, balancing the dataset. The performance of each predictor was reported using10-fold jackknife cross-validation data resampling technique. In the classifier development stage, diverse types of base-level predictors are trained in individual feature space. The preliminary decision acquired from base predictors is used to develop IDMS-HBC through the majority voting scheme. In this architecture, every predictor is supplied the same input dataset, in order that the final result of the combined predictor is achieved on the basis of the prediction of the individual predictors that was taken autonomously.

5.1.1 Development of Ensemble Classifiers

Here, a popular predictive models combination strategy of majority voting is adopted.

In this non-trainable strategy, a classification of an unlabeled example is carried out according to the predicted class that acquires the maximum number of votes. For example, suppose we have five different predictors, if a majority of the five predictors predicts a cancerous protein, then the prediction result of this protein sequence is taken as cancer.

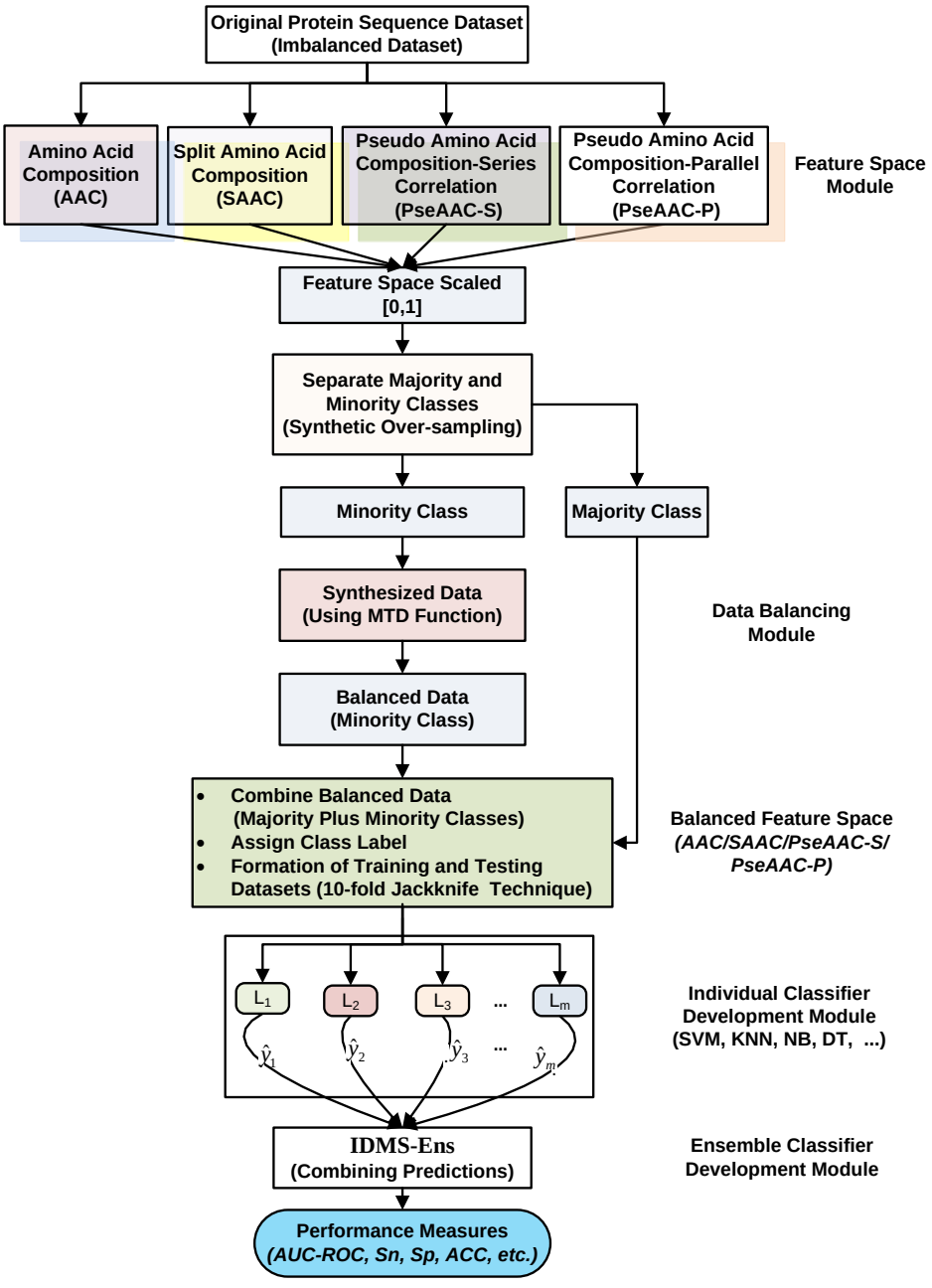


Figure 5.1 Basic block diagram of the proposed IDMS-HBC system.

Heterogeneous ensemble approach was developed with a set of m base-level predictors $\{C_1, C_2, \ldots, C_m\}$ using N samples in training dataset, $S_t = \{\mathbf{x}^{(n)}, t^{(n)}\}_{n=1}^{N}$, where

$\mathbf{x}^{(n)}$ represents the *n*th feature vector from feature spaces that are extracted using protein sequences correspond to target $t^{(n)}$. Using *j*th predictor on *i*th examples, predictions $\hat{y}_i^{\,j}$ i.e., $\hat{\mathbf{y}}_j^i = L_j(\mathbf{x}^i)$ is obtained. Each prediction belongs to one of the k classes, i.e., $\hat{\mathbf{y}}_j^i \in c^k$. This assigned *j*th label to one of the class from the set $\left\{c^1, c^2, ..., c^K\right\}$. The value of ensemble Z_{ens}^j is computed using majority-voting as:

$$Z_{ens}^j = \sum_j^m \Delta(\hat{y}_j^{\,i}, c^k), \qquad k = 1, 2, \ ..., \ \mathrm{K} \tag{5.1}$$

where, for our binary problem, K=2, and $\Delta(\hat{y}_j^{\,i}, c^k) = \begin{cases} 1 & \text{if } \hat{y}_j^{\,i} \in c^k \\ 0 & \text{otherwise} \end{cases}$. The data point x_j will be assigned to the class, which has maximum voting.

5.2 Results and Discussion

For the design of efficient ensemble system, several experiments are performed to explore the best combination of the base predictors. Table 5.1 shows the average diversity of predictors corresponding to each feature spaces. Lower the value of Q corresponds to higher the improvement in the proposed IDMS-HBC. For comparison, results obtained by implementing other well-known traditional ensemble approaches of GentleBoost, AdaBoostM1, and Bagging. The ensembles AdaBoostM1 and GentleBoost are based on decision tree classifiers. Whereas, Bagging ensemble is implemented using discriminant analysis based learning algorithm.

5.2.1 Performance of Individual Models

The performance of base predictors in terms of Acc, Sn, Sp, G-Mean, F-score, and MCC for C/NC and B/NBC are evaluated. The comparative results are shown in Table 5.2. The highest Acc and G-mean values corresponds to SVM model for PseAAC-S space using C/NC dataset.

5.2.2 Performance of the IDMS-HBC and Comparison with Other Approaches

The performance of the proposed system is provided in Fig. 5.2, Table 5.3 (first column), and Table 5.4 in different feature spaces for C/NC and B/NBC. Fig. 5.2 shows the improved ROC curves of the proposed ensemble system. From this figure, it is observed that IDMS-HBC using PseAAC-S and PseAAC-P spaces has provided

the best ROC curve among the individual feature space for C/NC and B/NBC, respectively. When combined predictions for all feature spaces, the average improvement of 0.23% and 0.20% are observed for C/NC and B/NBC datasets, respectively.

Table 5.1 Average Q statistic in different spaces for C/NC and B/NBC datasets.

Feature Space	Average Q Statistic	
	C/NC dataset	B/NBC dataset
AAC	0.0995	0.0995
SAAC	0.0994	0.0989
PseAAC-S	0.0993	0.0989
PseAAC-P	0.0996	0.0996

Table 5.2 Prediction performance of base predictors using different feature spaces for balanced datasets.

MTD-Model/ Feature space		C/NC dataset				B/NBC dataset			
		Acc	Sp	Sn	G_{mean}	Acc	Sp	Sn	G_{mean}
AAC	SVM	96.99	95.49	98.49	96.98	94.00	91.33	96.68	93.97
	KNN	93.12	93.30	92.95	93.12	92.61	93.58	91.65	92.61
	NB	90.75	82.08	99.42	90.34	93.84	88.97	98.72	93.72
	DT	92.08	91.45	92.72	92.08	92.34	91.76	92.93	92.34
	PNN	95.43	94.57	96.30	95.43	92.67	92.61	92.72	92.67
SAAC	SVM	96.65	95.26	98.04	96.64	94.11	88.22	100	93.93
	KNN	91.04	90.75	91.33	91.04	91.86	90.47	93.25	91.85
	NB	88.96	77.92	100	88.27	93.47	86.94	100	93.24
	DT	92.95	92.83	93.06	92.95	92.67	91.76	93.58	92.66
	PNN	94.05	95.38	92.72	94.04	92.72	93.90	91.54	92.71
PseAAC-S	SVM	97.05	97.99	96.14	97.06	94.06	90.90	97.22	94.00
	KNN	95.43	96.53	94.34	95.43	94.91	94.97	94.86	94.91
	NB	90.81	96.42	85.20	90.64	88.06	90.47	85.65	88.03
	DT	95.38	95.26	95.49	95.38	94.65	94.22	95.07	94.65
	PNN	95.49	98.96	92.02	95.43	94.06	95.18	92.93	94.05
PseAAC-P	SVM	96.19	93.18	99.19	96.14	94.33	89.83	98.82	94.22
	KNN	94.51	96.30	92.72	94.49	95.45	94.86	96.04	95.45
	NB	93.12	86.24	100	92.87	94.33	90.15	98.50	94.23
	DT	92.49	91.45	93.53	92.48	93.31	92.93	93.68	93.31
	PNN	96.18	98.61	93.75	96.15	95.13	97.43	92.83	95.10

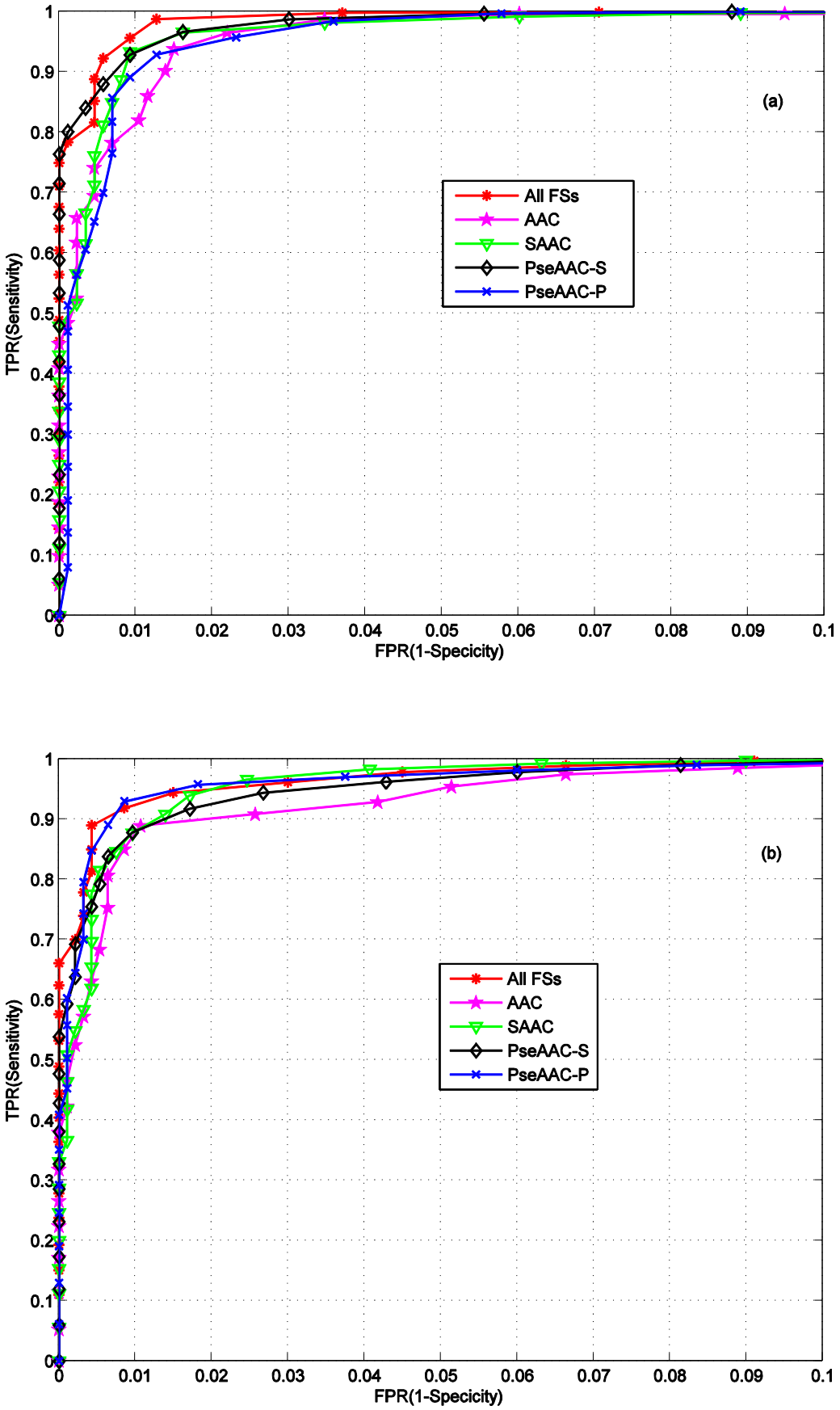


Figure 5.2 Partial ROC curves of the proposed IDMS-HBC for (a) C/NC and (b) B/NBC with balanced datasets.

Table 5.3 illustrates the performance comparison in terms of AUC of the proposed approach with other ensemble approaches of RF, AdaBoostM1 (Ada), Bagging (Bag), and GentleBoost (GnB) using balanced and original datasets. Now first we have comparison among conventional ensemble approaches. For balanced C/NC dataset, AdaBoostM1 has achieved improvement of 2.29%, 2.98%, 1.77%, and 5.13% in AUC using AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces, respectively. Whereas, for B/NBC dataset, AdaBoostM1 has shown enhancement of 19.15%, 10.61%, 4.95%, and 9.21% in AUC using AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces, respectively. As a whole, using balanced dataset, an average enhancement in the AUC of 2.16%, 7.13%, 7.10%, and 17.47% is observed for RF, AdaBoostM1, Bagging, and GentleBoost ensembles, respectively. It is inferred that balanced dataset with MTD is extremely beneficial for accurate performance of the predictors.

On the other hand, the proposed ensemble system has achieved the highest AUC of 99.79% and 99.58% using PseAAC-S and PseAAC-P feature spaces for C/NC and B/NBC balanced datasets, respectively. Overall, the IDMS-HBC ensemble learning system outperformed RF, AdaBoostM1, Bagging, and GentleBoost.

Table 5.3 Performance comparison in terms of AUC of the proposed IDMS-HBC with conventional ensemble approaches using balanced and original datasets.

Datasets/		Balanced					Original			
	Feature space	**Proposed Approach**	RF	Ada	Bag	GnB	RF	Ada	Bag	GnB
C/NC	AAC	99.54	99.24	97.02	97.97	97.50	97.53	87.73	91.54	89.94
	SAAC	99.61	99.67	98.14	98.32	98.88	97.32	95.16	92.73	96.28
	PseAAC-S	99.79	99.57	99.18	98.99	99.28	97.44	97.41	95.78	97.30
	PseAAC-P	99.51	99.21	98.20	98.59	98.46	96.80	93.07	93.96	93.40
B/NBC	AAC	99.22	99.07	95.99	98.22	96.89	95.46	76.84	89.23	76.52
	SAAC	99.51	98.97	97.13	98.72	98.00	96.83	86.52	90.36	90.89
	PseAAC-S	99.46	99.19	98.03	98.65	98.54	95.80	93.08	92.87	92.52
	PseAAC-P	99.58	99.19	98.02	98.71	97.94	92.64	88.81	91.52	89.83

In addition to AUC comparison, other measures of Acc, Sn, Sp, G-Mean, F-score, and MCC are used for evaluation of the proposed approach. In Table 5.4, the comparative results of the proposed and other ensemble approaches are shown using balanced datasets. For C/NC dataset, the highest Acc and G-mean values correspond to the proposed system using PseAAC-S space. For B/NBC dataset, the highest Acc

and G-mean values correspond to RF using PseAAC-P space, however, the proposed approach outperformed in terms of AUC. From Table 5.4, it is observed that overall the proposed system outperformed conventional approaches.

The well-known traditional ensemble approaches of RF, AdaBoostM1, Bagging, and GentleBoost have shown the poor performance compared to the proposed approach due to the following.

1) The previous traditional ensemble approaches have limited performance due to the small number of biological samples and class imbalance.
2) The previous approaches do not effectively combine the decisions of the base predictors using protein sequence features. However, in the proposed approach, we have used diverse learning algorithms and efficiently integrate their decisions to improve performance.
3) The previous approaches do not exploit effectively the useful diversity of base predictors. As a result, the performance of the previous approaches is poor.

In Table 5.5, a performance comparison is performed, in terms of AUC, of the proposed approach with previous approaches to examine which approach is adequate for the breast cancer problem. It is advantageous for the selection of data/classifier for future research of breast cancer. Delen et al. obtained the highest AUC of 84.9% and 76.9% for LR and DT, respectively [70]. Zhang et al. achieved maximum AUC of 87.4% and 93.2% for SVM and RF, respectively [69]. In another study, Khalilia et al. reported experimental results in terms of AUC for SVM (90.6%) predictor [71]. In recent study, Wang et al. attained maximum AUC of 84.2% and 82.0% for C_rLR and C_pDT approaches, respectively[68].

On the other hand, the proposed approach using Hd and Hb properties of amino acids based feature space together MTD has given the best performance in terms of AUC of 99.6% for breast cancer (Table 5.7) and 99.6% for cancer prediction (Table 5.4). Other ensemble approaches of AdaBoostM1, Bagging, and GentleBoost have provided AUC around 98.4% for breast cancer.

Usually, it is assumed that learning algorithms like RF, AdaBoostM1, and Bagging are performed better on imbalanced data. As, these approaches have the capability to improve the prediction performance by iteratively retraining the base predictors with a subset of most informative training data. Besides, we explore an

interesting aspect that presenting balanced data to these approaches would be adequately enhanced the performance (Tables 5.4 and 5.5). The proposed approach has achieved average improvement of 0.26%, 1.81%, 1.01%, and 1.34% in AUC over RF, AdaBoostM1, Bagging, and GentleBoost, respectively. On the other hand, the proposed system showed sufficient improvement over approaches from previous studies (Table 5.6). For example, the proposed approach showed higher results of 20.5%, 17.8%, 15.6%, 14.9%, and 12.4% in AUC over DT, C_pDT, C_rLR, LR, and SVM models, respectively.

Table 5.4 Performance comparison of the proposed IDMS-HBC with conventional ensemble approaches using balanced datasets.

Model/Feature space		C/NC dataset				B/NBC dataset			
		Acc	Sp	Sn	G_{mean}	Acc	Sp	Sn	G_{mean}
Proposed approach	AAC	97.40	97.40	95.72	97.38	95.66	98.18	93.15	95.63
	SAAC	96.82	98.96	94.68	96.80	96.57	98.93	94.22	96.55
	PseAAC-S	97.86	98.38	97.34	97.86	96.31	98.50	94.11	96.28
	PseAAC-P	97.23	98.84	95.61	97.21	96.95	97.54	96.36	96.95
RF	AAC	96.76	96.76	96.76	96.76	96.47	97.32	95.61	96.46
	SAAC	96.82	97.23	96.41	96.82	96.41	96.79	96.04	96.41
	PseAAC-S	97.72	98.12	97.07	97.59	96.84	97.64	96.04	96.83
	PseAAC-P	96.94	97.69	96.18	96.93	97.11	96.90	97.32	97.11
Adaboo-stM1	AAC	90.75	90.29	91.21	90.75	90.63	93.25	88.01	90.59
	SAAC	93.41	90.29	96.53	93.36	90.47	90.58	90.36	90.47
	PseAAC-S	96.36	95.26	97.46	96.35	91.92	93.15	90.69	91.91
	PseAAC-P	94.39	93.18	95.61	94.39	93.47	91.76	95.18	93.45
Bagging	AAC	93.41	90.29	96.53	93.36	93.95	92.51	95.40	93.94
	SAAC	94.51	91.33	97.69	94.46	94.43	90.04	98.82	94.33
	PseAAC-S	95.32	92.83	97.80	95.29	94.54	91.01	98.07	94.47
	PseAAC-P	94.22	90.52	97.92	94.15	94.75	91.22	98.29	94.69
GentleBoost	AAC	92.77	91.21	94.34	92.76	92.88	93.04	92.72	92.88
	SAAC	95.32	93.53	97.11	95.30	92.93	91.54	94.33	92.92
	PseAAC-S	96.07	95.49	96.65	96.07	93.74	91.76	95.72	93.72
	PseAAC-P	94.57	93.06	96.07	94.55	93.25	91.11	95.40	93.23

From the comparisons of the models it is verified that the MTD as preprocessor was effective to boost the forecast performance. It is found (Table 5.4) that although the *Sp* decreases slightly when applying MTD function, the *Sn* and *G*-

mean are improved. Thus, experimental results supported that the proposed system could be effectively used for cancer prediction in combination with MTD. On a whole, the experimental results demonstrated that the proposed heterogeneous system outperformed individual predictors, conventional ensembles, and previous approaches.

Table 5.5 Performance comparison of the proposed approach (IDMS-HBC) with previous approaches

Approach	AUC (%)	Acc	Sp	Sn	Dataset
LR	84.9	90.3	98.5	27.2	SEER data [70]
DT	79.3	89.6	98.8	21.4	SEER data [104]
SVM	87.4	NA	NA	NA	van de Vijver [105] and Wang [106] datasets [69]
RF	93.2	NA	NA	NA	-Do-
SVM	90.6	NA	NA	NA	NIS database [71]
C_rLR	84.2	75.1	75.0	76.2	SEER data [68]
C_pD	82.0	75.8	75.8	75.6	SEER data [68]
Proposed approach	99.8	97.86	98.38	97.34	Present study

The performance of the proposed approach is superior on three grounds. The first basis is the use of the most discriminative feature spaces, which are generated from physicochemical properties of amino acids. Second, the use of MTD as preprocessors for data balancing to improve accurate forecast performance. The third reason is that the proposed heterogeneous ensemble system efficiently combines the predictions of diverse types of learning algorithms at decision level.

This chapter has presented the performance of the heterogeneous ensemble system, IDMS-HBC that has shown considerable improvement, particularly, in PseAAC-S feature space. Overall, the proposed approach outperformed the individual, the conventional ensembles, and the previous approaches. In the next chapter, we shall discuss the performance of an intelligent ensemble system that combines specifically trained classifier on different feature spaces.

Chapter 6: Intelligent Ensemble Using Specific Classifier Trained on Different Feature Spaces

This chapter explains how high performance decision making ensemble classification systems are developed using diverse learning algorithms of RF, SVM, and KNN. Reliability and accuracy of cancer decision making ensemble systems mostly depend upon the discrimination power of feature extraction strategies. The variation of amino acid composition in cancer related proteins with respect to non-cancer proteins offers adequate discrimination for cancerous and healthy proteins and helpful for the development of prediction system. The '*IDM-PhyChm-Ens*' ensemble system is proposed based on the discrimination power of protein features and developed by combining the decision spaces of a specific classifier trained.

6.1 Variation of Amino Acid Composition in Cancer Proteins

Proteins present a rich source of information for the development of markers for diagnostics, prediction, and prognosis of breast cancer [107, 108]. The significance of such proteins is accentuated by the fact that entire anti-cancer drugs act on or through proteins [109]. It is observed from our analysis of variation of amino acid composition that it offers considerable potential for cancer prediction.

Traditionally in nutrition textbooks, amino acids are categorized into two types: essential amino acids and non-essential amino acids. Essential amino acids are obtained from our diet. They include: isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine. On the other hand, the body on its own manufacture nonessential amino acids. These amino acids are not necessarily obtained from diet. Nonessential amino acids consist of glutamate, alanine, aspartate, glutamine, arginine, proline, serine, tyrosine, cysteine, taurine, and glycine.

All of the native amino acids have similar general structure, except proline. The general structure of native amino acids holds a central α-carbon to an amino

group, a carboxylate group, a hydrogen atom, and an R side chain group. Several side chains are hydrophobic, while others are hydrophilic. On the contrary, in proline, amino group is secondary and is formed by ring closure between the R group and the amino nitrogen. In proline, rotation about carbon is impossible because of its rigidity to the peptide chain. This structural characteristic of proline is important in the structure and function of proteins with high proline content. Studies have revealed that proline plays a special role in cancer metabolism [110]. The proline biosynthetic pathway is considered a key mediator of breast tumor cell metastasis [111].

On the other hand, amino acids such as serine, threonine, tyrosine, asparagine, and glutamine have contained polar hydroxyl group and thus are called “hydrophilic” or “water-loving”. Polar hydroxyl group enables serine, threonine, tyrosine, asparagine, and glutamine to participate in hydrogen bonding, which is an important factor in protein structure. The hydroxyl groups serve other functions in proteins. Hydrogen-bonding capability of asparagine and glutamine has a significant effect on protein stability. Each protein has its own unique sequence of amino acids and native conformation. Any change in the sequence of amino acids, even one amino acid change, can potentially change the capability of protein to function. Consequently, study of variation of amino acid composition in cancer proteins with reference to non-cancer proteins is important. Because the Hd and Hb properties of amino acids are affected by the variation of amino acids in cancer proteins. Hence, such Hd and Hb properties of amino acids are quite useful in the prediction of cancer.

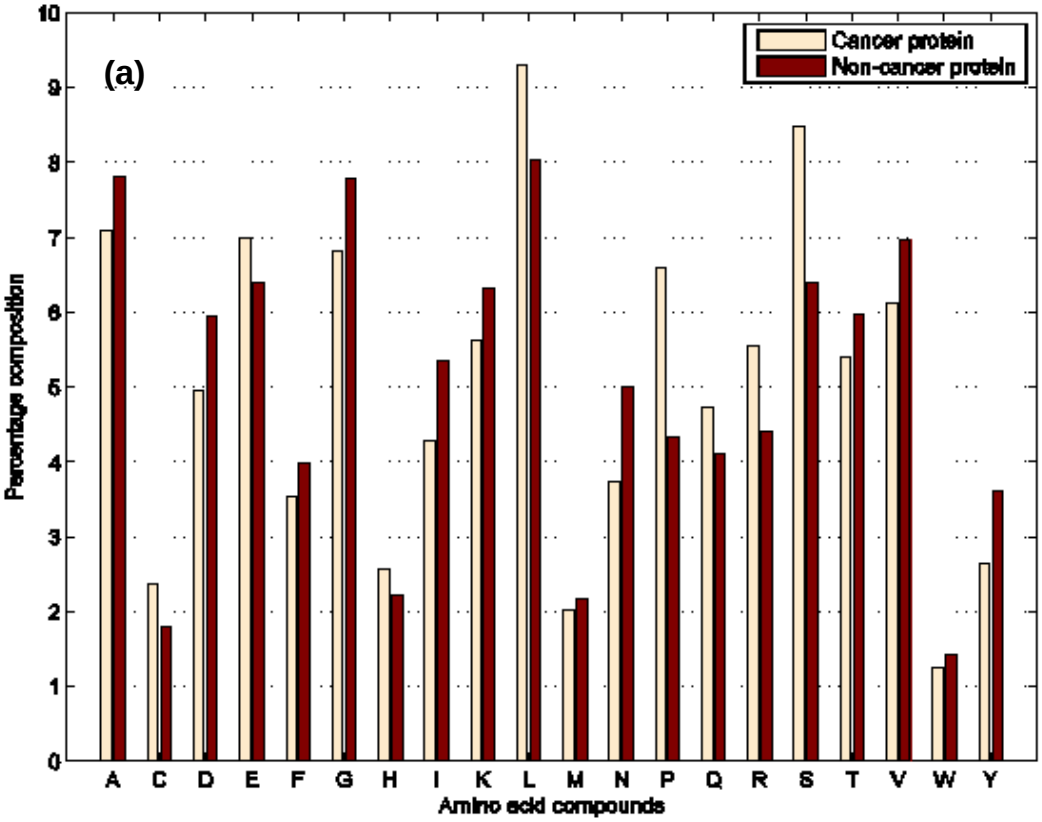
In Fig. 6.1(a-c), the percentage change is examined in concentration of amino acid compounds for general-cancer and breast-cancer proteins with respect to non-cancer protein. Fig. 6.1a and Fig. 6.1b show similar behavior for cancer and breast-cancer proteins sequences. The absolute percentage differences in all of the amino acids are relatively higher in breast-cancer proteins as compared to the general-cancer proteins, except C and S amino acids. From Fig. 6.1a, it is observed that amino acids Alanine (A), Aspartic acid (D), Phenylalanine (F), Glutamic acid (G), Isoleucine (I), Lysine (K), Asparagine (N), Threonine (T), Valine (V), and Tyrosine (Y) in non-cancer proteins are in excess than cancerous proteins. In contrast, Cysteine (C), Glutamic acid (E), Histidine (H), Leucine (L), Proline (P), Glutamine (Q), Arginine (R), and Serine (S) are more abundant (Fig. 6.1b) in breast cancerous proteins than non-cancer proteins.

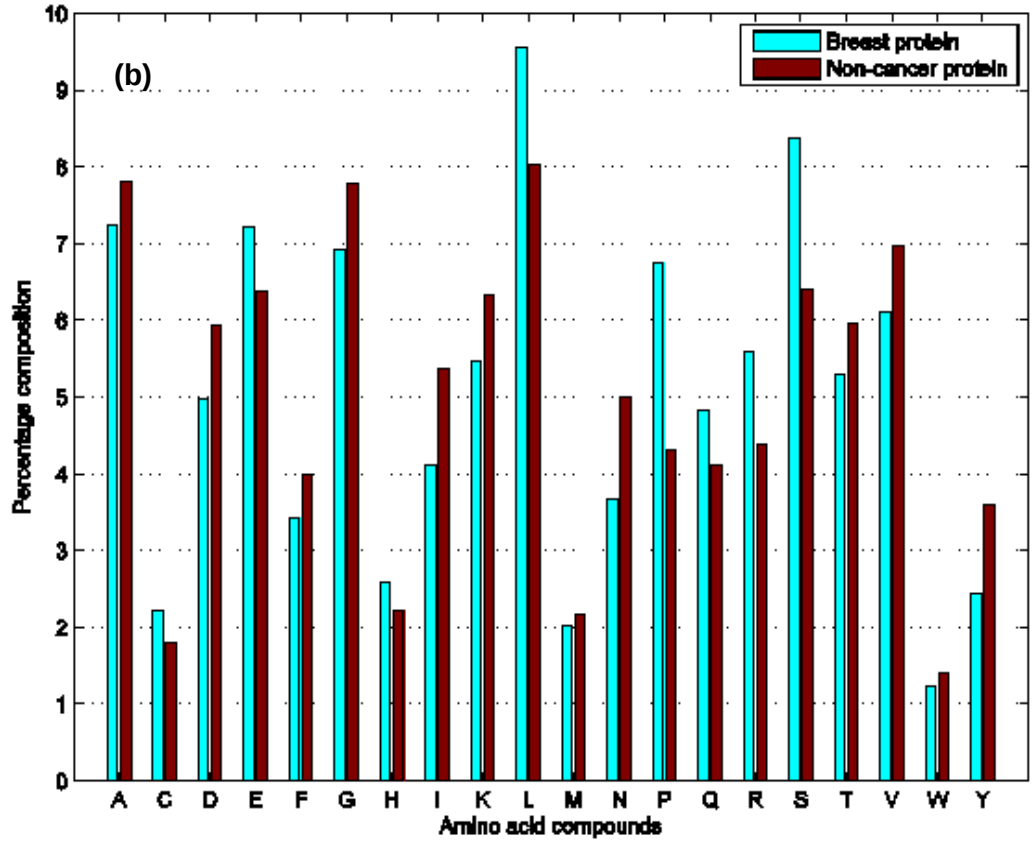
Further, Fig. 6.1c shows the behavior of amino acids compounds of cancerous and breast cancerous protein sequences with respect to non-cancerous proteins. The change in amino acids is relatively higher in breast-cancer proteins compared to the general-cancer proteins, except Cysteine and Serine amino acids. In this figure, values above the horizontal line correspond to the percentage increase in Cysteine, Glutamic acid, Histidine, Leucine, Proline, Glutamine, Arginine, and Serine for cancer related proteins. Values below this line indicate the percentage decrease in the following twelve amino acids: Alanine, Aspartic acid, Phenylalanine, Glutamic acid, Isoleucine, Lysine, Methionine (M), Asparagine, Threonine, Valine, Tryptophan (W), and Tyrosine for cancer related proteins with reference to non-cancer proteins. It is observed that Proline amino acid has maximum increase (~56%) followed by Serine (31%), Arginine (27.5%), Cysteine (23%), Leucine (19.2%), Glutamine (18%), Histidine (16.8%), and Glutamic acid (13.3%) in breast-cancer protein sequences. Tyrosine amino acid has revealed the maximum decrease (~32%) followed by Asparagine (27.1%), Isoleucine (23%), Aspartic acid (16%) and so on. It is noticed that higher change in the values of P, S, Y, C, R, and N offer high discrimination between breast cancerous and healthy proteins.

From Fig. 6.1, it is evident that the composition of all of the amino acids is altered in cancerous proteins. Based upon these results, it is expected that the variation composition of amino acid compounds in cancer proteins may help in the treatment of cancer and drug targeting. However, in the current work, the focus is to understand the role of such distribution of amino acid compounds in protein primary sequences using physicochemical properties in the early stages of cancer development.

The following text explains, how cancer proteins can be used to predict/diagnose cancer. How can a clinician use the information about the cancer related proteins to diagnose a patient?

Proteins of a tissue would reflect the initial changes caused by successive genetic mutations, which lead to cancer. Such changes/mutations would be exploited for the diagnosis of breast cancer. Fig. 6.1 illustrates the change in the composition of the amino acid molecules of the cancerous proteins with respect to non-cancerous proteins. It is observed that more disturbances occurred in P, S, Y, C, R, and N amino


(a)
Cancer protein
Non-cancer protein
Percentage composition
Amino acid compounds
A C D E F G H I K L M N P Q R S T V W Y


(b)
Breast protein
Non-cancer protein
Percentage composition
Amino acid compounds
A C D E F G H I K L M N P Q R S T V W Y

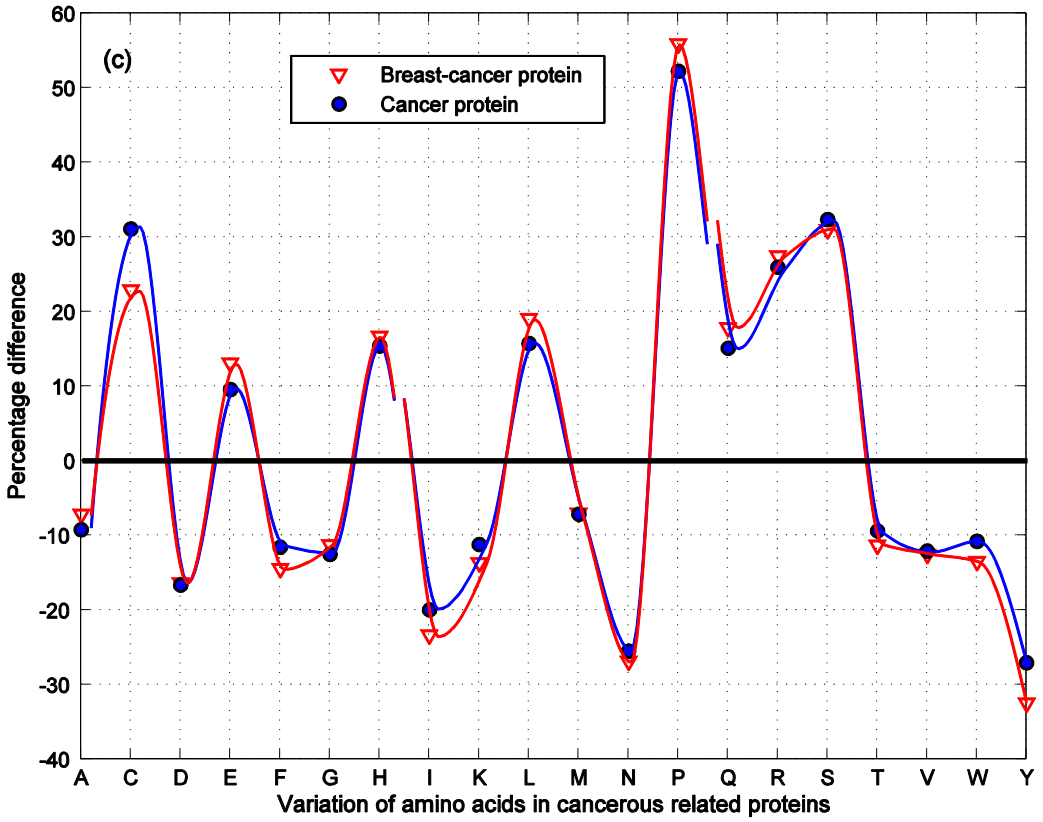


Figure 6.1 Variation of amino acid composition: (a) C/NC proteins, (b) B/NBC proteins, and (c) general-cancer and BC proteins with reference to NC protein sequences.

acid molecules. These molecules would offer high discrimination power between cancerous and non-cancerous proteins. In order to incorporate this discriminant feature, the molecular descriptors of amino acid sequences are formed using numerical values of their physiochemical properties of hydrophobiticty and hydrophicilty (see Table 2.3). These descriptors information are then transformed into various feature spaces using statistical and/or mathematical methods. During the training process, the proposed algorithm has extracted useful discriminant features (attributes).

6.2 The Proposed IDM-PhyChm-Ens System

Fig. 6.2 shows the basic block diagram of the proposed intelligent decision making system for the prediction of cancerous and healthy protein sequences. The proposed cancer prediction system *IDM-PhyChm-Ens* combines the decision spaces of a specific classifier trained on different feature spaces. It consists of three main modules: feature space generation, preprocessing, and classifier development modules. In preprocessing module, data balancing is performed using MTD function

to create diffuse samples for the minority class. This function oversamples the minority class in feature space.

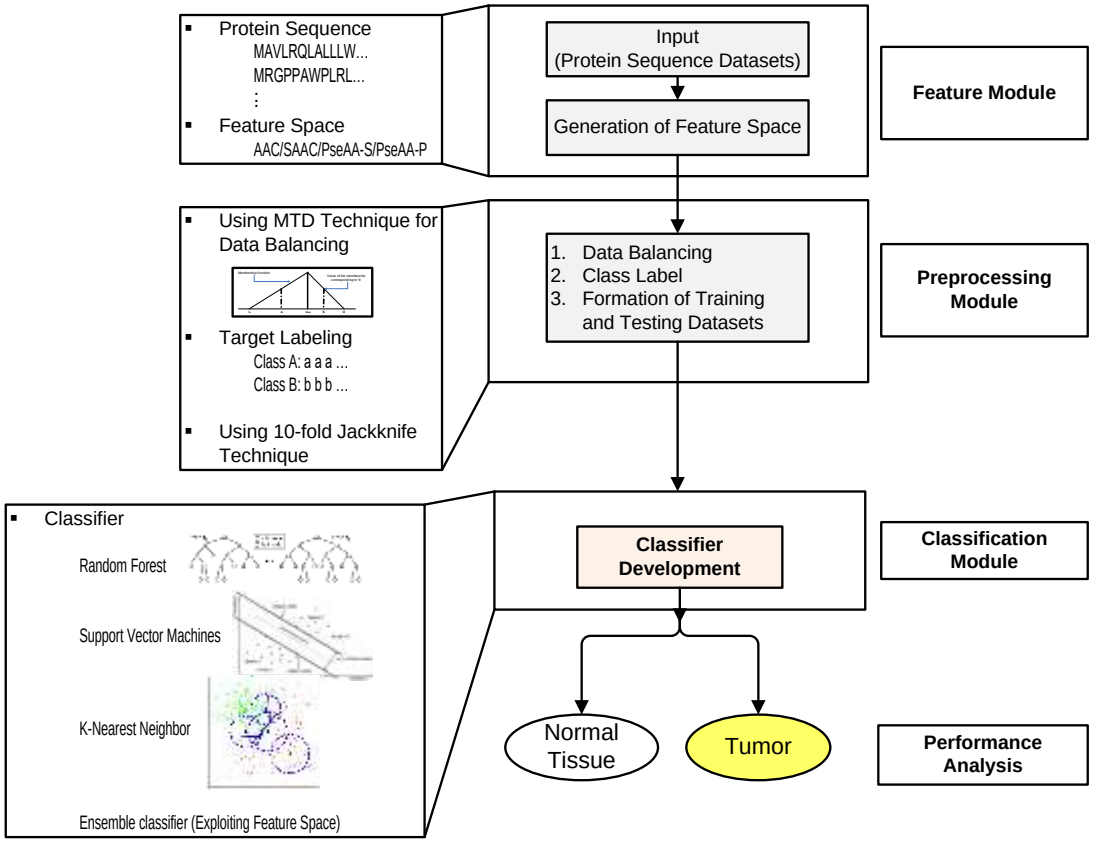


Figure 6.2 Basic block diagram of the proposed IDM-PhyChm-Ens prediction system.

6.2.1 Development of Ensemble Classifiers

Ensemble system may be formed using several diverse learning algorithms. Alternatively, ensemble system can be created with a single learning algorithm. In this study, homogeneous ensemble is developed with a specific base classifier, say, *L* using different feature spaces from the set $S = \{FS_1, FS_2, ..., FS_m\}$, where, $FS_1, FS_2, ..., FS_m$ are *m* different feature spaces.

A feature vector $\mathbf{V}_j^i$ of *i*th feature space, S_i, is formed form a protein data vector $\mathbf{x}_j$ as: $\mathbf{V}_j^i = FS_i(\mathbf{x}_j), \quad i = 1, 2, ..., m$. For a base classifier *C*, the predictions $\hat{y}_i^j$ are extracted as: $\hat{y}_i^j = L(\mathbf{V}_j^i)$. Each prediction $\hat{y}_i^j$ belongs to one of the two classes, i.e., $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_m\} \in \{c^1, c^2\}$. This assigned *j*th label to one of the class from

the set $\{c^1, c^2\}$. The value of ensemble Z_{ens}^j is computed using the majority-voting mechanism for all *m* feature spaces as:

$$Z_{ens}^j = \sum_i^m \Delta(\hat{y}_i^j, c^j), \qquad j = 1,2 \tag{6.1}$$

where, $\Delta(\hat{y}_i^j, c^j) = \begin{cases} 1 & \text{if } \hat{y}_i^j \in c^j \\ 0 & \text{otherwise} \end{cases}$. The data point x_j will be assigned to the class, which has maximum voting.

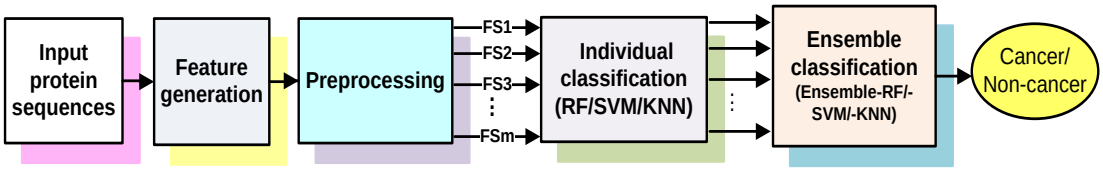


Figure 6.3 Framework of the proposed *IDM-PhyChm-Ens* ensemble scheme.

In this work, homogeneous ensembles by varying feature spaces of different dimensions are constructed. Fig. 6.3 demonstrates the various stages of the proposed ensemble system. Three algorithms RF, SVM, and KNN were selected as a single learning algorithm for the construction of homogeneous ensemble-RF, ensemble-SVM, and ensemble-KNN by exploiting different feature spaces. Each algorithm displays a different inductive bias and learning hypotheses such as instance-based, trees and statistics. Thus, each algorithm provides potentially more independent and diverse predictions.

The proposed system could be employed by academia, practitioners, or clinician, for the early diagnosis of breast cancer using protein sequences of the affected part of the organ. Protein sequences, which may be taken from DNA sequences, etc., could easily be supplied/presented to the system. If the specific order of amino acids in protein sequence is altered, the system will predict/diagnose cancer or else non-cancer.

6.3 Results and Discussion

Experimental results of the proposed *IDM-PhyChm-Ens* system are analyzed using: Performance Analysis (i) C/NC data, and (ii) B/NBC data. Performance of the proposed system is assessed in the individual and combined feature spaces using 10-fold cross-validation protocol. In the following subsections, performance analysis of

different ensemble classifiers with their counterpart individual classifier is presented.

Individual and Ensemble-RF Classifier

In this subsection, findings regarding individual and ensemble-RF classifiers using different feature spaces are discussed. Table 6.1 highlights the performance of RF based individual and ensemble classifiers using different feature spaces of AAC, SAAC, PseAAC-S, and PseAAC-P. It is observed that, for C/NC dataset, individual-RF has given the highest values of G_{mean} of 96.47% and F_{Score} of 96.64% for SAAC

Table 6.1 Performance of RF based individual and ensemble classifiers using different feature spaces.

Model/ dataset	Feature space	Acc	Sn	Sp	G_{mean}	F_{Score}	MCC
C/NC							
Individual-RF	AAC	95.78	99.19	91.56	95.30	95.55	72.01
	SAAC	96.76	99.88	93.18	96.47	96.64	72.00
	PseAAC-S	95.85	99.65	91.56	95.52	95.78	72.02
	PseAAC-P	95.61	99.54	91.21	95.28	95.56	72.00
Ensemble-RF	AAC+SAAC	99.08	98.84	99.31	99.07	99.07	70.22
	AAC+PseAAC-S	98.90	98.61	99.19	98.90	98.90	70.12
	AAC+PseAAC-P	98.90	98.50	99.31	98.90	98.90	70.04
	SAAC+PseAAC-S	99.48	99.54	99.42	99.48	99.48	70.57
	SAAC+PseAAC-P	99.25	99.42	99.08	99.25	99.25	70.56
	PseAAC-S+PseAAC-P	99.25	99.19	99.31	99.25	99.25	70.40
B/NBC							
Individual-RF	AAC	94.65	91.01	98.07	94.47	94.34	60.51
	SAAC	94.49	89.29	98.72	93.89	93.71	58.66
	PseAAC-S	95.13	93.90	96.57	95.23	95.17	63.84
	PseAAC-P	94.49	91.11	97.86	94.43	94.29	60.63
Ensemble-RF	AAC+SAAC	97.91	99.04	96.79	97.91	97.94	70.05
	AAC+PseAAC-S	97.16	99.36	94.97	97.14	97.22	70.77
	AAC+PseAAC-P	97.81	99.68	95.93	97.79	97.85	71.02
	SAAC+PseAAC-S	97.43	99.57	95.29	97.41	97.48	71.00
	SAAC+PseAAC-P	97.75	98.93	96.57	97.74	97.78	69.95
	PseAAC-S+PseAAC-P	97.11	99.68	94.54	97.08	97.18	71.27

feature space. However, for other feature spaces, individual-RF has provided the values of G_{mean} and F_{Score} near to 95.37% and 95.63%, respectively. However, ensemble-RF has yielded the highest value of 99.48% for Acc, G-mean, and F-score

using the combined feature space of SAAC+PseAAC-S. It is inferred that when the predictions of RF classifier using PseAAC-S feature space is combined with SAAC feature space, the performance of ensemble-RF is enhanced up to 3.01% in terms of G-mean. Therefore, for C/NC dataset, it is observed that SAAC feature space provides sufficient information for prediction.

Using B/NBC dataset, for PseAAC-S feature space, Table 6.1 indicates that individual-RF provided the highest values of Acc (95.24%), Sn (93.90%), Sp (96.57%), G_{mean} (95.23%), and F_{Score} (95.17%). However, the same feature space has the highest MCC value of 63.84%. Using combined feature space (AAC+SAAC), ensemble-RF classifier has given the highest values of Acc (97.91%), Sp (96.79%), G_{mean} (97.91%), and F_{Score} (97.94%) for breast cancer dataset. It is observed that when predictions of RF classifier using PseAAC-S feature space are combined with the predictions using PseAAC-P feature space, the performance of ensemble-RF classifier is enhanced by 7.43% for MCC measure. It is inferred that PseAAC-S feature space provides better information with the use of both individual-SVM and ensemble-SVM classifiers. Therefore, it can be concluded that, for the breast cancer prediction, PseAAC-S space using Hd and Hb properties of amino acids carries the most discriminant information from input data.

Individual and Ensemble-SVM Classifier

Table 6.2 reports the performance of SVM based on individual and ensemble classifiers using different feature spaces. For C/NC dataset, individual-SVM has provided the highest Acc (96.71%), G_{mean} (96.70%), and F_{Score} (96.73%) using PseAAC-S feature space. However, PseAAC-P feature space has provided maximum values of Sn (97.69%) and MCC (68.58%). In this table, using the combined feature space of PseAAC-S+PseAAC-P, Ensemble-SVM has given the highest Acc (97.63%), Sn (95.38%), Sp (99.88%), G_{mean} (97.60%), F_{Score} (97.58 %), and MCC (68.47%). However, when PseAAC-S feature space is combined at decision level with PseAAC-P, the performance of Ensemble-SVM is enhanced up to 0.90% (G_{mean}).

For B/NBC, Table 6.2 highlights that individual-SVM classifier has given the highest values of Acc (95.18%), Sn (93.04%), G_{mean} (95.16%), F_{Score} (95.08%), and MCC (62.80%) for PseAAC-S feature space. Using combined feature spaces of PseAAC-S and PseAAC-P, Ensemble-SVM has provided the highest values of Sp

(99.79%), and MCC (71.76%). However, ensemble-SVM has given the values of Acc (96.95%), Sp (94.22%), G_{mean} (96.91%) and F_{Score} (97.03%). It is observed that when predicated value of SVM classifier using PseAAC-S feature space is combined with predicated value of PseAAC-P feature space, at decision space, the performance of ensemble-SVM classifier is enhanced by 8.96% in the term of MCC. It is inferred that combined feature spaces of PseAAC-S and PseAAC-P have provided better information with the use of both individual-SVM and ensemble-SVM classifiers. Thus it is concluded that these feature spaces possess the most discriminant information using Hd and Hb properties of amino acids for the prediction of breast cancer.

Table 6.2 Performance of SVM based individual and ensemble classifiers using different feature spaces.

Model/ dataset	Feature space	Acc	Sn	Sp	G_{mean}	F_{Score}	MCC
C/NC							
Individual-SVM	AAC	95.72	96.88	94.57	95.72	95.77	67.6
	SAAC	95.72	96.76	94.68	95.72	95.77	67.49
	PseAAC-S	96.71	97.57	95.84	96.70	96.73	68.35
	PseAAC-P	96.47	97.69	95.26	96.47	96.52	68.58
Ensemble-SVM	AAC+SAAC	96.71	93.76	99.65	96.66	96.61	67.80
	AAC+PseAAC-S	97.11	94.57	99.65	97.08	97.03	68.14
	AAC+PseAAC-P	97.23	94.68	99.77	97.19	97.15	68.18
	SAAC+PseAAC-S	97.11	94.45	99.77	97.07	97.03	68.08
	SAAC+PseAAC-P	97.23	94.57	99.88	97.19	97.15	68.12
	PseAAC-S+PseAAC-P	97.63	95.38	99.88	97.60	97.58	68.47
B/NBC							
Individual-SVM	AAC	94.59	92.72	96.47	94.57	94.49	62.50
	SAAC	95.07	91.97	91.97	95.02	94.92	61.55
	PseAAC-S	95.18	93.04	97.32	95.16	95.08	62.80
	PseAAC-P	94.97	91.97	97.97	94.92	94.81	61.57
Ensemble-SVM	AAC+SAAC	96.57	99.68	93.47	96.52	96.68	71.46
	AAC+PseAAC-S	95.77	99.68	91.86	95.69	95.93	71.75
	AAC+PseAAC-P	96.15	99.25	93.04	96.10	96.26	70.96
	SAAC+PseAAC-S	96.31	99.57	93.04	96.25	96.42	71.39
	SAAC+PseAAC-P	96.95	99.68	94.22	96.91	97.03	71.33
	PseAAC-S+PseAAC-P	96.20	99.79	92.61	96.13	96.33	71.76

Individual and Ensemble-KNN Classifier

Table 6.3 shows the performance of KNN based individual and ensemble classifiers using different feature spaces. For C/NC dataset, it is observed that individual-KNN has provided the highest Acc (96.01%), Sn (95.14), G_{mean} (96.01%), and F_{Score} (95.98%) using PseAAC-S feature space. However, individual-KNN, for other feature spaces, have given Acc, Sn, G-mean and F-score nearly 94.74%, 93.83%, 94.73% and 94.69%, respectively. From this table, ensemble-KNN using combined feature spaces AAC and SAAC has provided the highest values of 98.84%, 99.07%, 99.07% and 70.22% for Sn, G-mean, F-score, and MCC, respectively.

Using B/NBC dataset, Table 6.3 indicates that individual-KNN classifier has yielded the highest Acc (94.59%), Sp (94.86%), G_{mean} (94.59) %, and F_{Score} (94.58%) for PseAAC-P feature space. However, for PseAAC-S feature space, Sn and MCC measures have the highest values of 94.43% and 64.64%, respectively. Ensemble-KNN, for AAC+SAAC features space, has provided the highest values of Acc (94.54%), Sp (89.72%), G_{mean} (94.42%), and F_{Score} (94.79%). However, for PseAAC-S+PseAAC-P feature spaces, ensemble-KNN has given the highest values for Sn 99.89% and MCC 73.19%. It is noted that decision of Ensemble-KNN is enhanced by 8.55% (MCC), when the predicted values of KNN classifier in PseAAC-S feature space is combined with the predicted values in PseAAC-P feature space. The combined feature spaces of PseAAC hold the most useful information for B/NBC dataset.

6.3.1 Overall Performance Comparison

Figs. 6.4-6.6 highlighted the overall performance comparison of individual and ensemble classifiers using the most informative feature spaces in terms of G-mean and F-score. For C/NC dataset, Fig. 6.4 shows that individual-SVM in PseAAC-S feature space and ensemble-RF in SAAC+PseAAC-S feature spaces have performed better. For breast cancer, Fig. 6.5 highlights that individual and ensemble of RF classifiers have performed better in PseAAC-S and AAC+SAAC feature spaces, respectively. Fig. 6.6 demonstrates the overall performance comparison of ensemble classifiers for the highest performing feature spaces in terms of MCC measure. For C/NC dataset, from Fig. 6.6, it is observed that ensemble-RF performs better than other ensemble classifiers using SAAC+PseAAC-S feature spaces. However, for

B/NBC dataset, ensemble-KNN is better than other ensemble classifiers using combined feature spaces of AAC+SAAC. Therefore, the prediction system

Table 6.3 Performance of KNN based individual and ensemble classifiers using different feature spaces.

Model/ dataset	Feature space	Acc	Sn	Sp	G_{mean}	F_{Score}	MCC
C/NC							
Individual-KNN	AAC	94.74	93.29	96.18	94.73	94.66	63.18
	SAAC	93.99	94.34	93.64	93.99	94.01	64.62
	PseAAC-S	96.01	95.14	96.88	96.01	95.98	65.28
	PseAAC-P	95.49	93.87	97.11	95.48	95.42	63.77
Ensemble-KNN	AAC+SAAC	93.87	98.84	99.31	99.07	99.07	70.22
	AAC+PseAAC-S	94.22	88.67	99.77	94.06	93.88	65.94
	AAC+PseAAC-P	93.76	87.51	100.0	93.55	93.34	65.59
	SAAC+PseAAC-S	94.68	89.71	99.65	94.55	94.40	66.28
	SAAC+PseAAC-P	94.10	88.44	99.77	93.93	93.75	65.87
	PseAAC-S+PseAAC-P	94.57	89.36	99.77	94.42	94.27	66.16
B/NBC							
Individual-KNN	AAC	93.36	93.58	93.15	93.36	93.38	63.74
	SAAC	92.56	92.61	92.51	92.56	92.56	62.63
	PseAAC-S	94.54	94.43	94.65	94.54	94.53	64.64
	PseAAC-P	94.59	94.33	94.86	94.59	94.58	64.50
Ensemble-KNN	AAC+SAAC	92.88	99.89	85.87	92.61	93.35	73.19
	AAC+PseAAC-S	94.00	99.79	88.22	93.83	94.33	72.58
	AAC+PseAAC-P	94.06	99.68	88.44	93.89	94.37	72.38
	SAAC+PseAAC-S	93.68	99.68	87.69	93.49	94.04	72.52
	SAAC+PseAAC-P	93.79	99.68	87.90	93.61	94.14	72.48
	PseAAC-S+PseAAC-P	94.54	99.36	89.72	94.42	94.79	71.69

developed using RF algorithm has shown good performance for the prediction of breast cancer. From Fig. 6.5 and Fig. 6.6, it is observed that using the combined feature space of AAC+SAAC, ensemble-RF (in terms of G-mean and F-score) and ensemble-KNN (in terms of MCC) have provided better performance for B/NBC dataset. Further, it is observed that ensemble-RF have performed well in SAAC+PseAAC-S and AAC+SAAC feature spaces to predict C/NC and B/NBC datasets, respectively.

6.3.2 Comparison with Previous Studies

In Table 6.4, a performance comparison of the proposed IDM-PhyChm-Ens approach with previous approaches is carried out to analyze which approach is better for breast

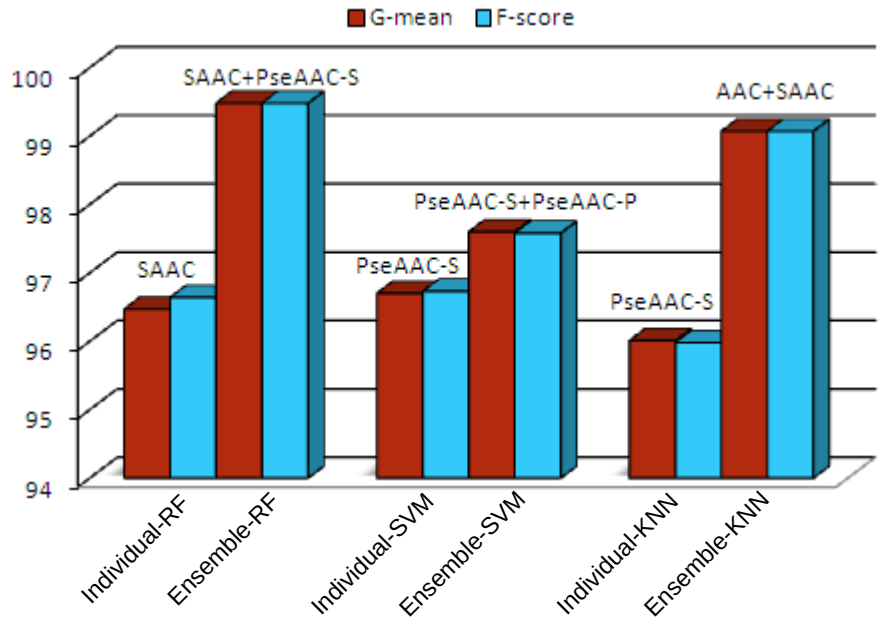


Figure 6.4 Comparison of individual and ensemble classifiers based on RF, SVM, and KNN algorithms using the highest performing feature spaces in terms of G-mean and F-score for C/NC dataset.

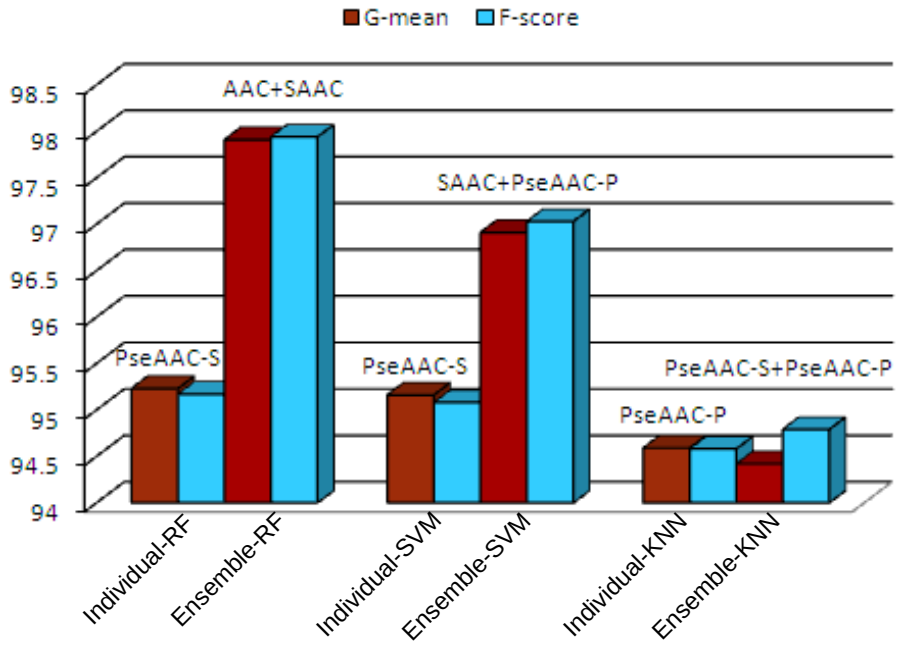


Figure 6.5 Comparison of individual and ensemble classifiers based on RF, SVM, and KNN algorithms using the highest performing feature spaces in terms of G-mean and F-score for B/NBC dataset.

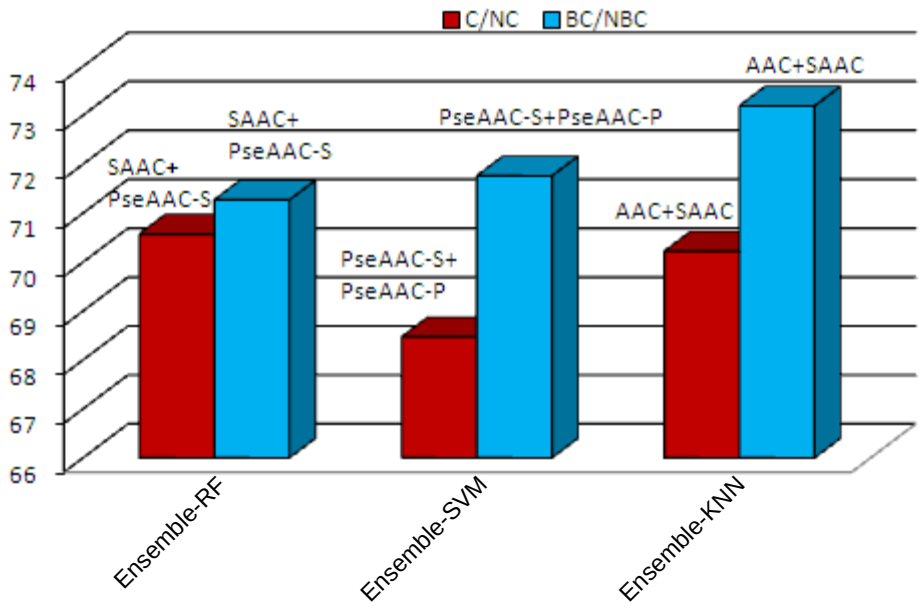


Figure 6.6 Comparison of ensemble classifiers using the highest performing feature spaces in terms of MCC.

cancer problem using different features and classifiers. It could be helpful for the selection of features and classifiers for future research of breast cancer. Multiple sonographic and textural features based classifiers have provided an accuracy in the range 83.80-86.92% [57]. On the other hand, mammographic features using SVM classifier enhanced accuracy up to 93.73% [112]. Socio-demographic and other cancer specific information based features have prediction accuracy in range 93.60-94.70% [52].

Topological indices based QPDR models have maximum accuracy of 90.5 % [12]. Clinical features computed from a digitized image of FNA of a breast mass were also used for breast cancer prediction along with various ML approaches [62, 113]. Optimized-LVQ, Big-LVQ, AIRS, and LDA models provided accuracy near to 96.80% [113]. The prediction performance using clinical features enhanced the accuracy up to 97.40% for Fuzzy-GA, AR +NN, and SVM+EAs approaches.

On the other hand, our ensemble-RF system using Hd and Hb properties of amino acids based combined feature space SAAC+PseAAC-S has given the best prediction accuracy of 99.48% for C/NC. Ensemble-RF system using combined feature space AAC+SAAC has given the best prediction accuracy of 97.91% for B/NBC. Similarly, ensemble-SVM using feature spaces PseAAC-S+PseAAC-P and SAAC+PseAAC-P based on Hd and Hb properties of amino acids has yielded an

improved accuracy of 97.63% and 96.95% than previous SVM based approaches [52, 57, 63, 112, 114, 115]. Ensemble-NB, using Hd and Hb properties of amino acids, based in SAAC+PseAAC-S feature space has yielded the best prediction accuracy of 98.32% for C/NC dataset. Ensemble-NB in AAC+SAAC feature space has given the

Table 6.4 Comparison of the prediction accuracies achieved from the proposed prediction system IDM-PhyChm-Ens with other classifiers from literature.

Model	Feature extraction strategy		Acc (B/NBC) (%)		Dataset
KNN [57]	Sonographic and textural features		83.80		Harbin digital ultrasound image database
ANN [57]			86.60		-Do-
SVM [57]			86.92		-Do-
ANN [70]	Incidence and population based features		91.20		SEER data
Decision Tree [70]			93.60		-Do-
Decision Tree [52]	Socio-demographic and cancer specific information based features		93.62		Iranian ICBC dataset
ANN [52]			94.70		-Do-
SVM [52]			95.70		-Do-
SVM [112]	Mammographic features		93.73		UCI repository: Mammographic Mass
Decision Tree [54]	Clinical features (10 features for each cell nucleus computed form a digitized image of a fine needle aspirate (FNA) of a breast mass)		94.74		UCI repository: Wisconsin-breast-cancer
Optimized-LVQ [62]			96.70		-Do-
Big-LVQ [62]			96.80		-Do-
AIRS [62]			97.20		-Do-
LDA [113]			96.80		-Do-
SVM [115]			97.20		-Do-
Fuzzy-GA [116]			97.36		-Do-
AR +NN [53]			97.40		-Do-
SVM+EAs [63]			97.07		-Do-
Fuzzy-SVM [114]	Clinical features extracted with Principal Component Analysis		96.35		-Do-
Ensemble (NF KNN QC) [117]	Information gain based selected clinical features		97.14		-Do-
	C/NC	B/NBC	C/NC	B/NBC	
QPDR [12]	*pTIe* (embedded)	*TIe+ dTIe* (embedded)	90.0	91.80	Same as present study
Ensemble-RF	SAAC+PseAAC-S	AAC+SAAC	99.48	97.91	**Present study**
Ensemble-SVM	PseAAC-S+PseAAC-P	SAAC+PseAAC-P	97.63	96.95	-Do-
Ensemble-KNN	SAAC+PseAAC-S	PseAAC-S+PseAAC-P	94.68	94.54	-Do-
Ensemble-NB	SAAC+PseAAC-S	AAC+SAAC	98.32	98.88	-Do-

best prediction accuracy of 98.88% for BC dataset. Therefore, ensemble-RF and ensemble-NB yielded the almost same level of performance for cancer datasets. This analysis has shown that NB and RF approaches based prediction systems are worth for the prediction of cancerous protein sequences problem.

The previous approaches have attained the poor performance compared to the proposed approach due to the following.

1) The previous approaches are developed by employing different types of features, which could not provided sufficient information for improved performance.

2) The previous approaches do not effectively combine the decisions of the base predictors. However, the proposed approach is developed using diverse learning algorithms, which are trained on different feature spaces. Decision spaces of the specific classifier are effectively combine to enhanced the performance.

The performance of IDM-PhyChm-Ens system is superior due to two reasons. The first reason is the use of descriptors derived from physicochemical properties of amino acids in protein primary sequences. These descriptors have a potential to accommodate the variation of amino acid composition in cancer and breast-cancer protein sequences with reference to non-cancer proteins. The second reason is the use of ensemble classifier that combines the predictions of a specific learning algorithm at decision level using different feature spaces.

In this chapter, the development of ensemble system using a specific learning algorithm of RF, SVM, and KNN trained on different feature spaces has been discussed. It is observed that these ensemble systems are more effective than their individual counterparts. In the next chapter, another heterogeneous ensemble system using classifier stacking GP-based evolutionary approach is developed for the prediction of human breast cancer.

Chapter 7: Classifier Stacking Based Evolutionary Ensemble System

In this chapter, a classifier stacking GP-based evolutionary ensemble system is developed, further exploiting the variation of amino acid molecules associated with breast cancer. GP evolution process combines the diverse-type of useful information of base predictors by generating better decision space than conventional ensemble approaches. This evolutionary ensemble system is expected to provide ameliorated performance by taking advantage of its better exploration and exploitation ability in search space. Comparative analysis of the proposed system with conventional ensemble approaches of AdaBoostM1, Bagging, GentleBoost, and Random Subspace is discussed.

7.1 The Proposed Can-Evo-Ens System

The proposed novel Can-Evo-Ens (Cancer Evolutionary Ensemble) system has two level classifier structures; in 1st level NB, KNN, SVM, and RF are used as base-level classifiers for generation of preliminary predictions and, then in 2nd level GP is employed to develop meta-classifier for the fusion of base classifiers. The 1st level base classifiers are trained on the original input dataset, and their predicted results are extracted called meta-data. Then, 2nd level meta-classifier is trained on this new dataset to obtain final prediction. The most suitable threshold value of the best evolved predictor is computed using the Particle Swarm Optimization (PSO) technique. Fig. 7.1 shows the basic block diagram of the proposed Can-Evo-Ens system. Fig. 7.1a indicates data preprocessing, development and predictions stacking of base-level predictors (1st level). Fig. 7.1b demonstrates the working principle of GP (2nd level), PSO based optimal threshold, and system performance evaluation module. GP evolution process has combined effectively the diverse-type of useful information of base predictors by generating better decision space than individual and conventional ensemble approaches.

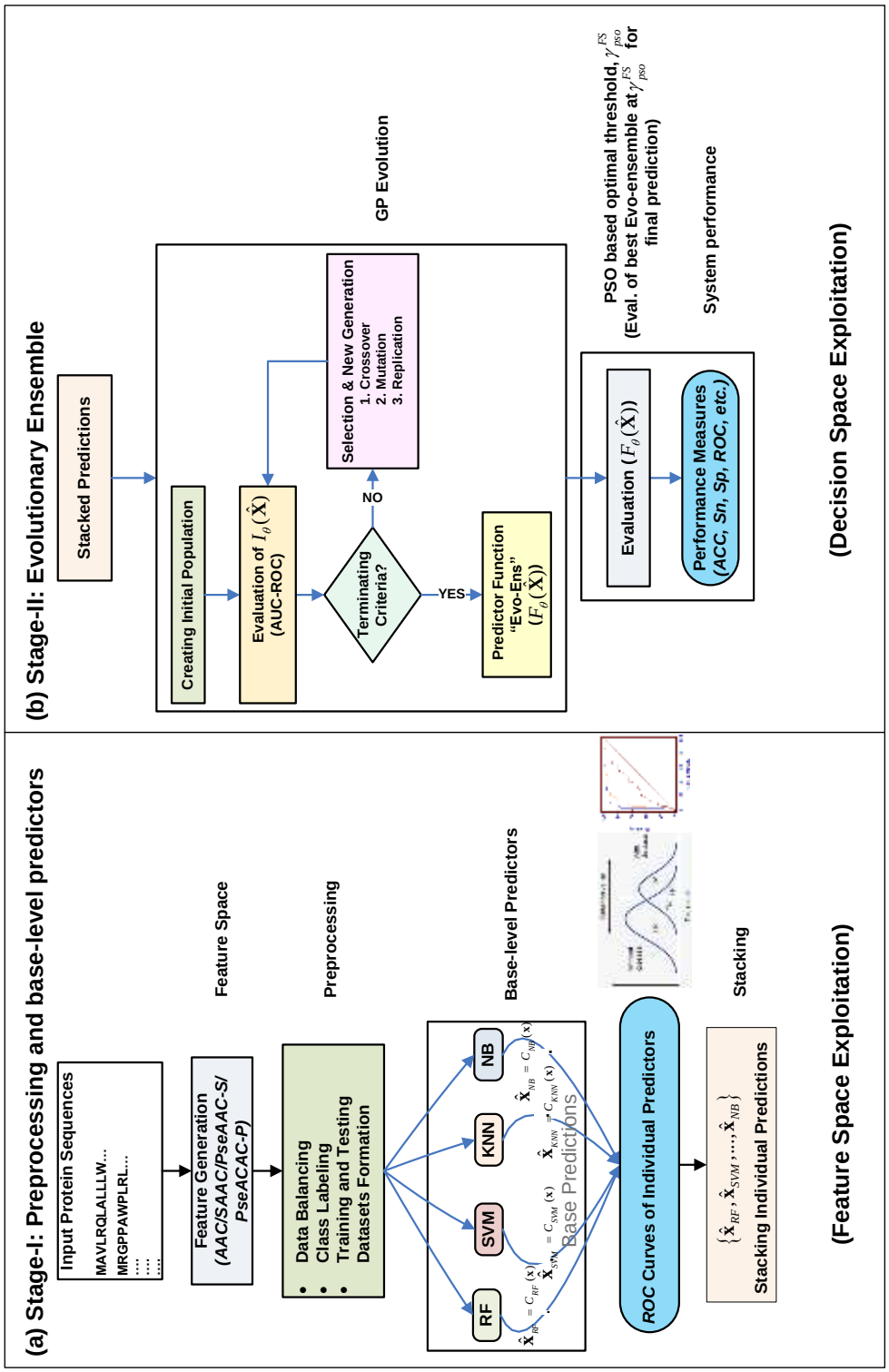


Figure 7.1 Basic block diagram of the proposed "Can-Evo-Ens" ensemble system: (a) Stage-I represents the data preprocessing and base-level predictors, (b) Stage-II indicates GP evolution process.

7.1.1 Data Preprocessing

In preprocessing module, various tasks of data balancing, class labeling, and training/testing datasets formation are performed. The dataset of protein amino acid

sequences is randomly divided into two separate parts: (i) training dataset (*Trn*) and (ii) validation or testing dataset (*Tst*). On the average, about $(1-e^{-1}) \approx 2/3$ data is used for training and about $e^{-1} \approx 1/3$ data is leaving for model evaluation. Each base predictor (classifier) is train for *Trn* dataset and thereby obtained a new meta-data (*Trn_pred*) for the development of GP based evolutionary ensemble "Evo-Ens". The performance of base level prediction models is reported using *Trn* dataset. On the other hand, independent validation dataset (*Tst*) is used to evaluate the performance of Can-Evo-Ens system (i.e. the optimal model generated using PSO to find the optimal threshold). Here, the "Evo-Ens" represents the output of the GP module that is developed at the end of the GP process. However, "Can-Evo-Ens" denotes the complete cancer evolutionary ensemble system.

7.1.2 Classifier Stacking

The predictions of base-level predictors are stacked to develop Evo-Ens model. A set of base-level predictors $\{C_1, C_2,\ldots,C_m\}$ is constructed on training dataset of *N* samples, $S_t = \left\{\left(\mathbf{x}^{(n)}, t^{(n)}\right)\right\}_{n=1}^{N}$, where $\mathbf{x}^{(n)}$ represents the *n*th feature vector of protein sequences correspond to target $t^{(n)}$. We obtained a set of predictions $\left\{\hat{\mathrm{x}}_1^i, \hat{\mathrm{x}}_2^i, \ldots, \hat{\mathrm{x}}_m^i\right\}$ in numerical form using *m* base predictors i.e., $\hat{\mathrm{x}}_j^i = C_j(\mathbf{x}^i)$ represents *j*th predictor on *i*th examples. Note that, these predictions in the form of numerical values are very important related to the evolved GP function because the base prediction values are used as the leaf nodes of individuals in GP. The most commonly used arithmetic, trigonometric, logarithmic, etc. functions are defined in GPLAB and it assured and automatically care that no leaf node becomes a negative value. Hence, at meta level training, m-dimensional feature vector is formed $\hat{\mathbf{X}}_m = (\hat{\mathrm{x}}_1, \hat{\mathrm{x}}_2, \ldots, \hat{\mathrm{x}}_m)$. In this way, for GP training a new meta data of N sample points is created, i.e., $S_d = \left\{(\hat{\mathbf{X}}^{(n)}, t^{(n)})\right\}_{n=1}^{N}$. GP technique maps prediction vectors $\hat{\mathbf{X}}^{(n)}$ of base-level predictors to target labels $t^{(n)}$. At the end of GP process, the best numerical Evo-Ens function, represented $F_\theta(\hat{\mathbf{X}})$, is developed.

7.1.3 GP Evolution Process

GP evolutionary approach can effectively exploit the search space to find the best candidate solution [85]. In GP evolution process, first, initial population of predictor functions is constructed, i.e., $\phi = I_\theta(\hat{\mathbf{X}})$, where $\hat{\mathbf{X}} \in \Re^m$, $\phi \in \Re$, $\Re^m$ is m dimensional real vector and Θ denotes set of selected GP parameters. A set of functions (plus, sin, log, etc.), variables (x, y, etc.), and randomly generated constants is provided to find a suitable structure of target function. The candidate solutions $I_\theta(\hat{\mathbf{X}})$ are representing in the form of tree structure. This tree-like representation consists of variable size. Adaptable tree representation automatically discovers the underlying useful pattern within data. The terminal set of tree comprises of useful feature vectors and random constants created with uniform distribution. The most informative values of parameters and variables are chosen. The initial population of 100 individuals is generated using *ramped half-and-half* method.

In *second step*, fitness scores of individual candidates $I_\theta(\hat{\mathbf{X}})$ are evaluated. The fitness score demonstrates how well GP individual moves toward the optimal solution. The success of evolutionary approach depends upon the accurate design of the fitness function. In this study, AUC-ROC is used as GP-fitness criterion. The area is calculated using the trapezoids method, defined as:

$$I_s = \sum_{k}^{N_t - 1} \frac{1}{2}(FP_{k+1} - FP_k)(TP_{k+1} - TP_k) \tag{7.1}$$

where, N_t denotes the number of class thresholds, and TP_k and FP_k represent the true positive and the false positive at class threshold *k*. The equation sums the area of the individual trapezoids of heights h_1 and h_2 and width *w*, fitted under the ROC points. This measure returns values between 0 and 1; higher the value, better the performance. The AUC corresponds to the probability that a minority class example is correctly predicted. Fig. 7.2 represents computation of GP solution on input examples and ROC curve for each threshold *T*. The maximum fitness score I_s indicates how successfully an individual $I_\theta(\hat{\mathbf{X}})$ moves towards the optimal point.

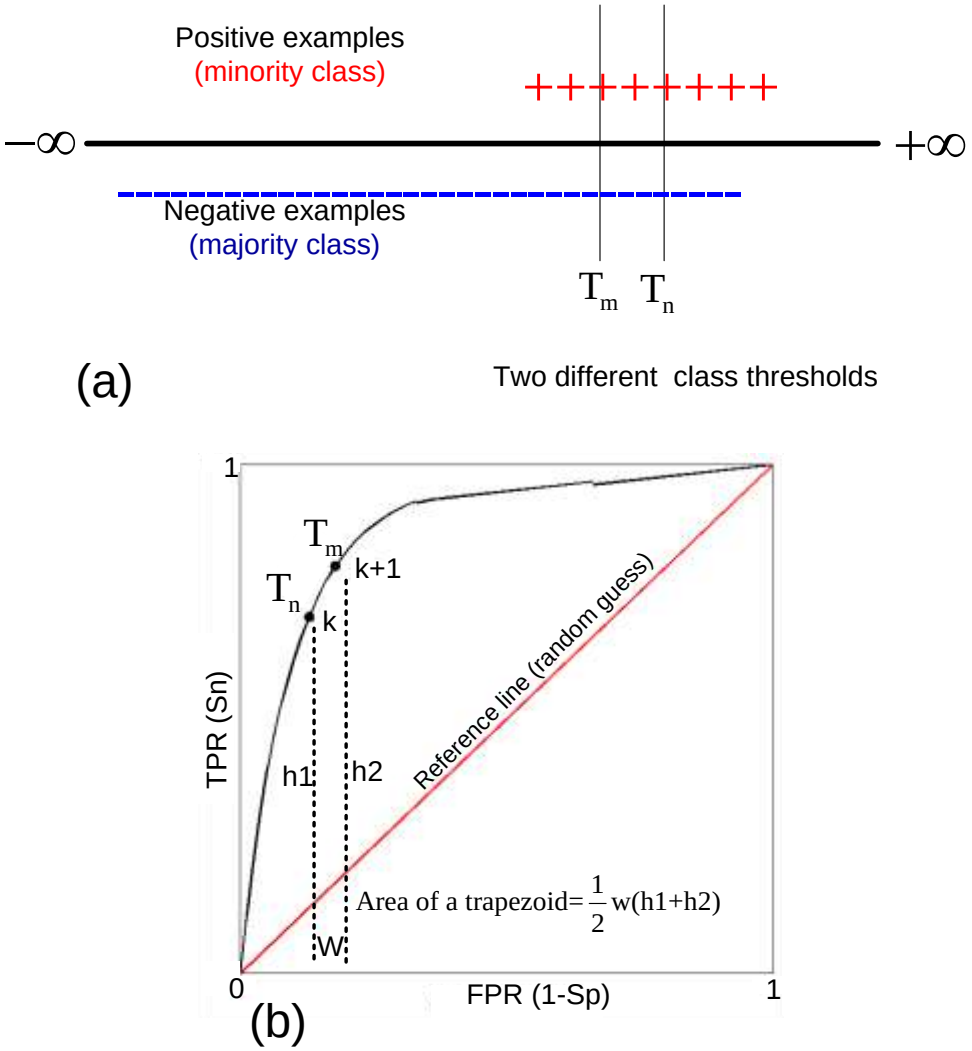


Figure 7.2 (a) GP solution in the form of numeric outputs, where '+' and '-' be the cancer and non-cancer output classes, respectively, and T_m and T_n be the two different class thresholds; (b) ROC curve, which indicates two thresholds points T_m and T_n and corresponding area of a trapezoid.

During *third step*, the current population is used to choose the best candidates. In population of size P_s, the fitness probability of individual candidates, $I_\theta(\hat{\mathbf{X}})$ is obtained as

$$\Pr(I_\theta) = \frac{I_\theta}{\sum_{P_s} I_\theta} \tag{7.2}$$

where, $\sum_{P_s} I_\theta$ denotes the total fitness of the total population size. The individual with higher probability values has grater possibility to produce offspring. *Fourth step* creates new population by applying the crossover, mutation, and replication operators to individual parents. GP process is initialed to automatically speed up the

convergence process while maintaining the population diversity. During simulation, the genetic operator probabilities, crossover, and mutation rates are selected to '*variable*'. The GPLAB software adopts the rates of these search operators to reflect their performance, based on the procedure as described in [118]. The rate of the operator will increase that has performed well in the previous search processes. That operator has more chance to produce offspring again. Two simulation stopping criteria are used:

(i) maximum generations reach up to 200, or

(ii) maximum fitness score $(I_s \geq 0.999)$.

Ultimately, the best individual $F_\theta(\hat{\mathbf{X}})$ in the population, i.e., $I_\theta(\hat{\mathbf{X}}) \rightarrow F_\theta(\hat{\mathbf{X}})$ is chosen. Since values of $F_\theta(\hat{\mathbf{X}})$ varies with different threshold, it is desirable to choose the most suitable threshold to classify the dataset as cancer and non-cancer.

7.1.4 Computing Optimal Threshold

Conventional search techniques such as the grid search can be used to obtain the threshold values for Evo-Ens functions. However, for these conventional search techniques, we have to adjust manually suitable grid range and step size. The computational complexity of the problem depends on the grid range and its step size. However, for efficient computation, it is preferred to use the PSO based intelligence technique to find the optimal threshold for the Evo-Ens functions in different feature spaces. The dimensionality of the search space is the same as the dimension of the feature space. In PSO, initial population is started with a random set of threshold (particle) for the GP expression. The position of each particle refers to a candidate solution. PSO finds the fitness value of each particle to determine personal best (Pbest) and global best (Gbest) particles. The particles are moved toward the optimal area by updating their position and the velocity according to the algorithm [119]. PSO has selected the best threshold value γ_{pso}^{FS} for each feature space (*FS*). The best predictions $\hat{g}_{Ens}^{FS}$ are computed for Cancer and Non-Cancer classes as:

$$\hat{g}_{Ens}^{FS} = \begin{cases} Cancer, & \text{if } F_\theta^{FS}(\hat{\mathbf{X}}) \geq \gamma_{pso}^{FS} \\ Non-Cancer, & \text{otherwise} \end{cases} \tag{7.3}$$

7.2 Parameter Settings

In the following subsections, first, we explain selection procedures of the parameter setting of individual predictors and then describe how Evo-Ens is develop using the proper parameter setting in different feature spaces. Several simulations were carried out to select these parameters. A summary of the parameters used for the development of individual predictors (base level classifiers) and Evo-Ens are provided in the sections 7.2.1 and 7.2.2, respectively.

7.2.1 Parameter Settings of Individual Predictors

The individual predictors are trained using *Trn* datasets in AAC, SAAC, PseAAC-S, and PseAAC-P feature spaces. In the design of Evo-Ens, four diverse types of base classifiers NB, KNN, SVM, and RF are selected. Here, RF classifier is selected, instead of Decision Tree, as a base classifier, because Decision Tree classifier has low accuracy and higher variance. On the other hand, Bayesian approach is computationally less complex and this approach has proved to be very effective in biological data classification problem. Table 7.1 provides the description of parameter selection of different predictors. This table gives the selected values of *K* for different feature spaces. It shows the optimal parameter values of SVM predictors of the error term *C* and width of the Gaussian kernel σ. Similarly, for RF predictor, Table 7.1 indicates the number of trees (*ntree*) and number of selected variables (*mtry*) in various feature spaces.

7.2.2 Parameter Setting of Evo-Ens

Table 7.1 shows the summary of the necessary parameters to generate the best GP expressions. Several GP simulations were carried out to adjust these parameters using GPLAB3 toolbox in MATLAB R2013a environment. During training phase, approximately 15-20 runs of GP were performed. The computational time of each GP simulation depends on various input parameters of input data size, population size, and number of generations. While, developing $F_{\theta}^{FS}(\hat{\mathbf{X}})$ function, the most instructive arithmetic and trigonometric functions are chosen during GP evolution. Its performance depends on the useful combination of local information of the protein amino acid sequence.

For C/NC dataset, Fig. 7.3a shows the growth of the best individual generations using various feature extraction strategies. A considerable improvement in fitness of the best GP individual is observed up to 8 generations for PseAAC-S and 50 generations for PseAAC-P feature spaces. For AAC and SAAC spaces, enough improvement is found in the best GP individual up to 40 generations. After about 50 generations, there occur smooth fitness transition and the best-fit individual converges

Table 7.1 Summary of the parameters settings for individual predictors and the proposed Evo-Ens.

Algo.	Parameter	Description/Value
KNN	K=Number of Nearest Neighbors	*C/NC dataset*: K=1 for AAC and SAAC, and K=22 for PseAAC-S and PseAAC-P *B/NBC dataset*: K=5 for AAC and SAAC, and K=22 for PseAAC-S, and PseAAC-P
SVM	(C, σ), where C denotes penalty parameter of the error term and σ be the width of the Gaussian kernel	*C/NC dataset*: (120, 0.01) for AAC, (100, 0.05) for SAAC, PseAAC-S and (100, 0.005) for PseAAC-P *B/NBC dataset*: (100, 0.00045) for AAC and SAAC, and (100, 0.0005) for PseAAC-S and PseAAC-P
RF	Number of trees to generate (*ntree*), and number of candidate predictors (*mtry*)	*C/NC dataset*: *ntree* =60 for AAC, *ntree* =90 for SAAC, *ntree* =40 for PseAAC-S, and *ntree* =160 for PseAAC-P *B/NBC dataset*: *ntree* =60 for AAC, *ntree* =30 for SAAC, *ntree* =40 for PseAAC-S, and *ntree* =30 for PseAAC-P *mtry* $\approx \sqrt{n}$, where *n* is dimension of feature space.
GP	Terminals/non-terminal set	Set of feature vectors, $\hat{\mathbf{X}} = \{X_1\ X_2\ X_3\ X_4\}$,where X_1, X_2, X_3, and X_4 represent the predictions of base-level predictors RF, SVM, KNN, and NB, respectively. Parameters set, $\theta = \{\text{Functions, Constants}\}$ Functions = $\{\text{plus, minus, times, divide, log, sin, cos, exp, power}\}$ Constants = random numbers from interval [0, 1]
	Fitness criterion	Area under the receiver operating characteristic curve (AUC-ROC)
	Population Initialization	Ramped half and half
	Selection method	Generational
	Generations	180 for C/NC dataset and 140 for B/NBC dataset
	Population size	85
	Expected offspring	rank89
	Sampling	Tournament
	Maximum Tree depth/size	32
	Survival criterion	Keep the best individual
	Operators probabilities	Select crossover and mutation rates to 'variable'

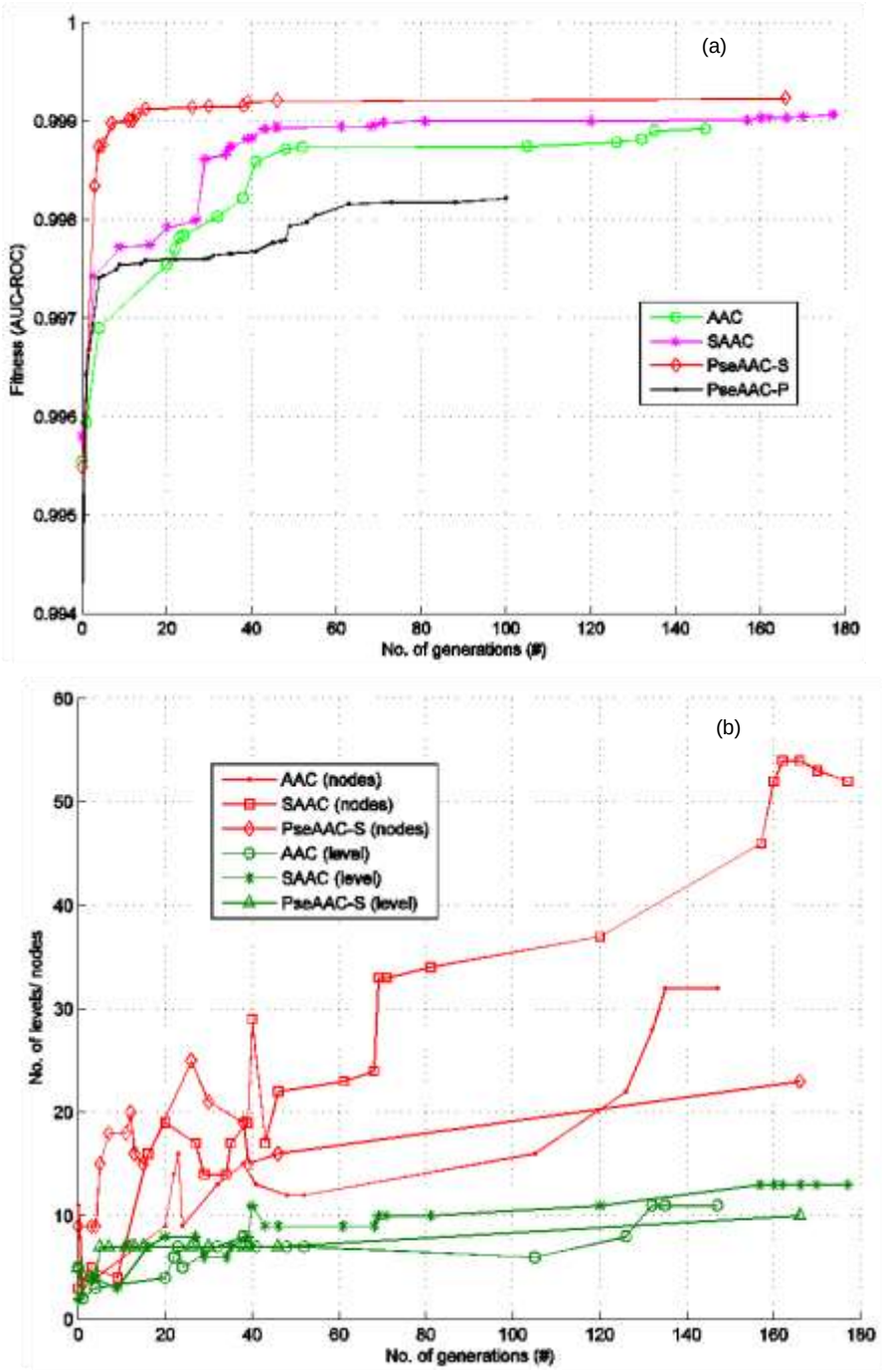


Figure 7.3 For cancer dataset, complexity of the best GP individual in each generation with respect to (a) fitness criterion (b) number of nodes and level of tree depth against the number of generations

to near optimal for AAC, SAAC, and PseAAC-P. It is noticed that after 18 generations, best-fit individual for PseAAC-S space converges to optimal or near optimal. Fig. 7.3b shows the complexity of the best individual in each generation against the number of generations for cancer dataset. This figure highlights that complexity is raised as the number of generations increased. Generally, in GP

evolution process, constructive blocks are produced that try to shield useful building blocks. Consequently, in many regions of Fig. 7.3b, the size of GP individual grows with no adequate enhancement in the performance curve of the best candidate. Under the bloating phenomenon, trees keep growing without corresponding improvements in fitness that is some branches do not play a part in the performance [120]. Thus, the total number of nodes increases exponentially. Particularly for SAAC space, the average tree depth becomes very large. It is observed that PseAAC-S has average nodes and tree depth relatively small as compared to other feature extraction strategies.

During training phase, several simulation runs were carried out and the best GP function is used for reporting of results. For C/NC dataset, the best numerical functions are developed using AAC, SAAC, PseAAC-S, and PseAAC-P spaces. These functions in prefix form are given below:

$F_\theta^{AAC}(\hat{\mathbf{X}})$ =minus(minus(minus(minus(X_1 , X_3),times(X_4 ,sin(X_2))),minus(times (minus(minus(X_1 , X_3),log(cos(sin(plus(plus(X_1 , X_3),sin(X_2))))))),cos(divide(sin (X_2),0.63288))), X_4)), X_3) (7.4)

$F_\theta^{SAAC}(\hat{\mathbf{X}})$ =divide(times(minus(X_3 ,times(times(plus(log(plus(X_3 , X_1)), X_2),times (cos(sin(divide(minus(abs(divide(X_4 , X_4)),minus(X_3 ,minus(0.59321, X_2))),minus (X_1 ,0.68599)))),sin(X_2))),times(X_4 ,sin(divide(times(X_2 ,cos(X_3)),plus(sin (times(cos(X_3), X_3)),sin(X_1)))))), X_3),log(log(cos(X_2)))) (7.5)

$F_\theta^{PseAAC-S}(\hat{\mathbf{X}})$ =minus(sin(divide(X_3 , X_1)),minus(plus(sin(X_1),minus(cos(minus(X_4 ,times (log(log(X_1)),0.49498))),times(X_3 ,log(0.016173)))), X_4)) (7.6)

and

$F_\theta^{PseAAC-P}(\hat{\mathbf{X}})$ =plus(plus(cos(X_2),times(times(minus(minus(minus(cos(cos(minus (minus(minus(X_4 , X_3), X_3), X_3))),cos(sin(X_1))), X_1), X_3),abs(sin(cos(minus (minus(cos(X_2), X_2),cos(minus(abs(X_3), X_3)))))), X_3)),cos(minus(minus(cos(X_2), X_2),cos(minus(abs(X_3), X_3))))) (7.7)

The tree structures of Evo-Ens for SAAC and PseAAC-P are shown in the Fig. 7.4(a&b). These graphical representations demonstrated the functional dependency of

expressions on the predictions of base predictors $\hat{\mathbf{X}} = \{X_1\ X_2\ X_3\ X_4\}$ along with some selected arithmetic and trigonometric functions.

For breast cancer dataset, Fig. 7.5a shows the improvement of the best-fit individual in different feature spaces. Fig. 7.5a demonstrates that in later generations sufficient fitness improvement is observed. For PseAAC-S and PseAAC-P spaces, after 30 generations a smooth fitness transition take place and the best-fit individual converges to near optimal. For AAC and SAAC spaces, the best GP individual is improved up to 65 generations.

Fig. 7.5b shows increase in complexity in terms of level of tree depth and number of nodes against the number of generations. Generally, in the GP evolution process, constructive blocks are formed that attempt to protect the useful genetic material. Consequently, in many regions of this figure, the size of best GP individuals grows with no adequate improvement in the performance curve of the best individual. Under the bloating phenomenon, trees keep growing without any improvements in fitness that is some branches do not contribute in the performance [120]. Thus, the total number of nodes increases exponentially. It is observed that for PseAAC-S space, average tree depth becomes large and best genome's total number of nodes increases as compared to other feature spaces. However, PseAAC-P space has a relatively small average number of nodes and tree depth as compared to other feature spaces. It is inferred that as the genome complexity increases, the performance curve of the best individual approaches towards the optimal solution.

For breast cancer, optimal numerical functions $F_\theta^{FS}(\hat{\mathbf{X}})$ are developed for AAC, SAAC, PseAAC-S, and PseAAC-P spaces. These functions are given below:

$F_\theta^{AAC}(\hat{\mathbf{X}})$ =minus(cos(times(X_1 ,times(minus(cos(X_2), X_2),minus(cos(X_1),0.75846)))), X_3) (7.8)

$F_\theta^{SAAC}(\hat{\mathbf{X}})$ =plus(minus(plus(minus(X_1 , X_3),plus(minus(plus(X_3 ,plus(minus(cos (divide(minus(X_1 ,sin(cos(X_1))), X_1)), X_1), X_3)), X_3),minus(0.49256, X_3))), X_3),minus(minus(minus(cos(X_3), X_3),minus(X_3 ,minus(minus(sin(X_1), X_3), X_1))), X_1)) (7.9)

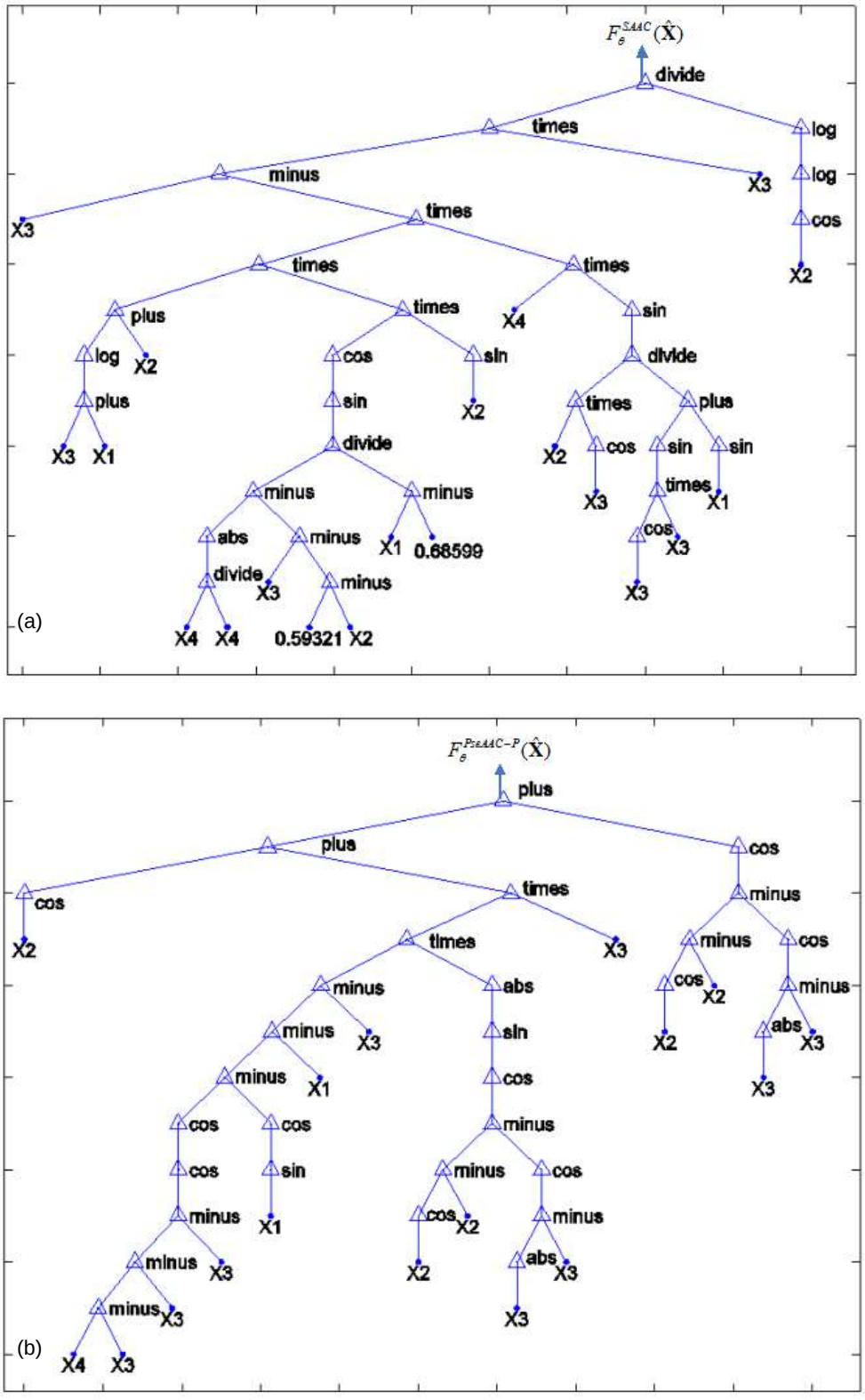


Figure 7.4 Trees of the best individual of Evo-Ens predictor using (a) SAAC and (b) PseAAC-P spaces for C/NC dataset.

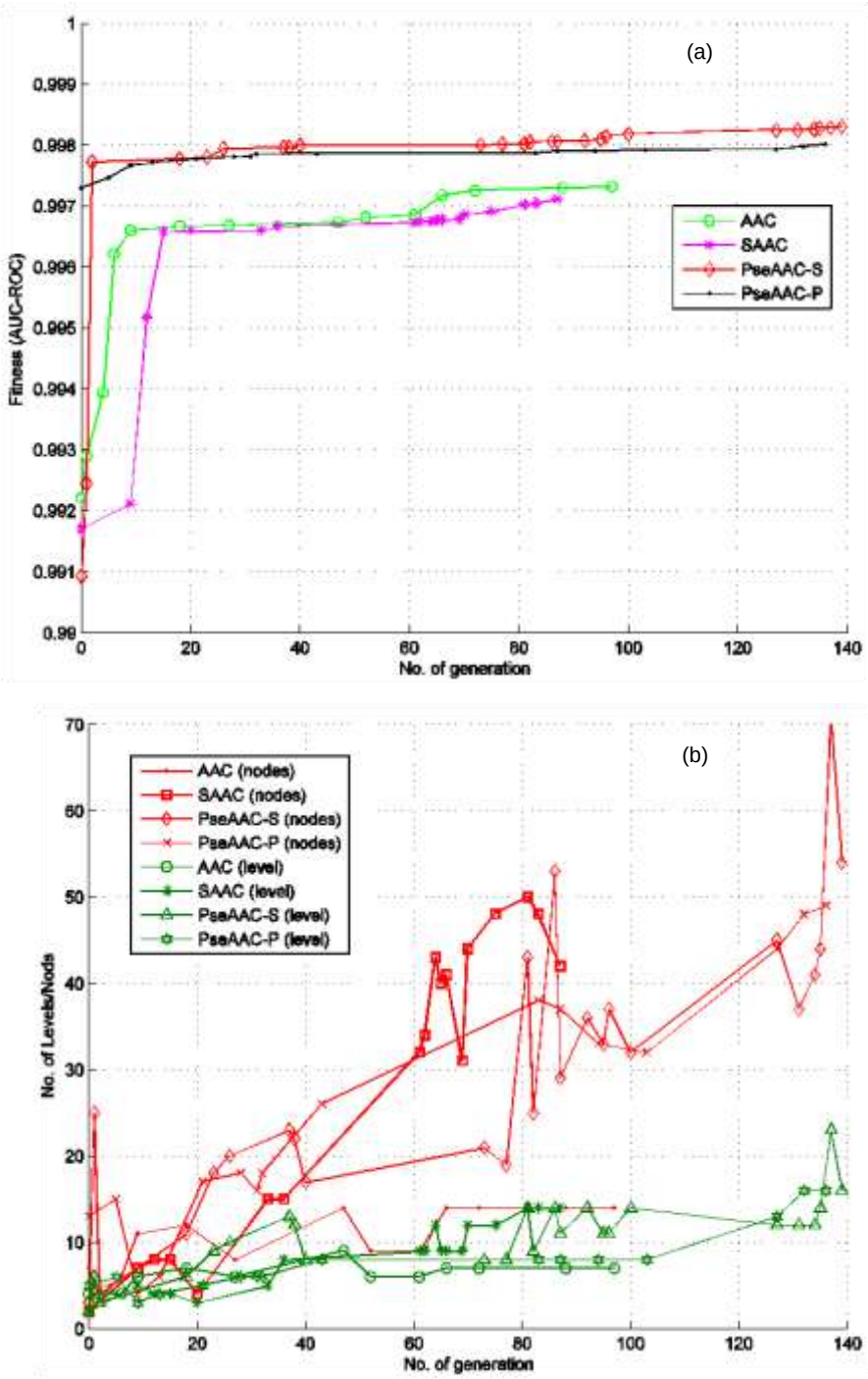


Figure 7.5 For breast cancer dataset, (a) improvement in best GP individuals in each generation, (b) increase in complexity with respect to number of nodes and level of tree depth against generations.

$F_{\theta}^{PseAAC-S}(\hat{\mathbf{X}})$=minus(sin(sin(sin(sin(cos(plus(times(X_2 ,cos(times(sin(X_2),sin(abs(X_4))))),times(minus(X_2 , X_4),times(plus(times(X_3 ,minus(X_1 ,times(plus(plus(X_4 ,0.0057079),minus(0.70858, X_1)),cos(X_1)))), X_3),plus(X_4 ,0.33607))))))))),minus (plus(X_1 , X_3),cos(times(X_2 ,times(X_2 ,plus(X_4 ,minus(X_3 , X_4))))))) (7.10)

and

$F_\theta^{PseAAC-P}(\hat{\mathbf{X}})$ =minus(cos(cos(minus(plus(plus(cos(minus(cos(X_1),divide(times (divide(X_3 ,cos(cos(log(times(X_4 , X_4)))))),cos(cos(0.30436))),cos(0.49206)))), X_3), X_3),plus(X_3 ,sin(X_1)))))),divide(times(divide(cos(0.14211),cos(0.14211)),divide (times(cos(0.30436), X_3), cos(times(X_1 , X_1)))),cos(0.49206))) (7.11)

Generally, the complex structure of predictor functions may not be analyzed / understood by human expert. However, these empirical functions can be easily computed by machine. For example, Fig. 7.6 show the tree structures of the best individual functions for PseAAC-P spaces. These graphical representations highlight the functional dependency of GP expressions on predictions of base predictors.

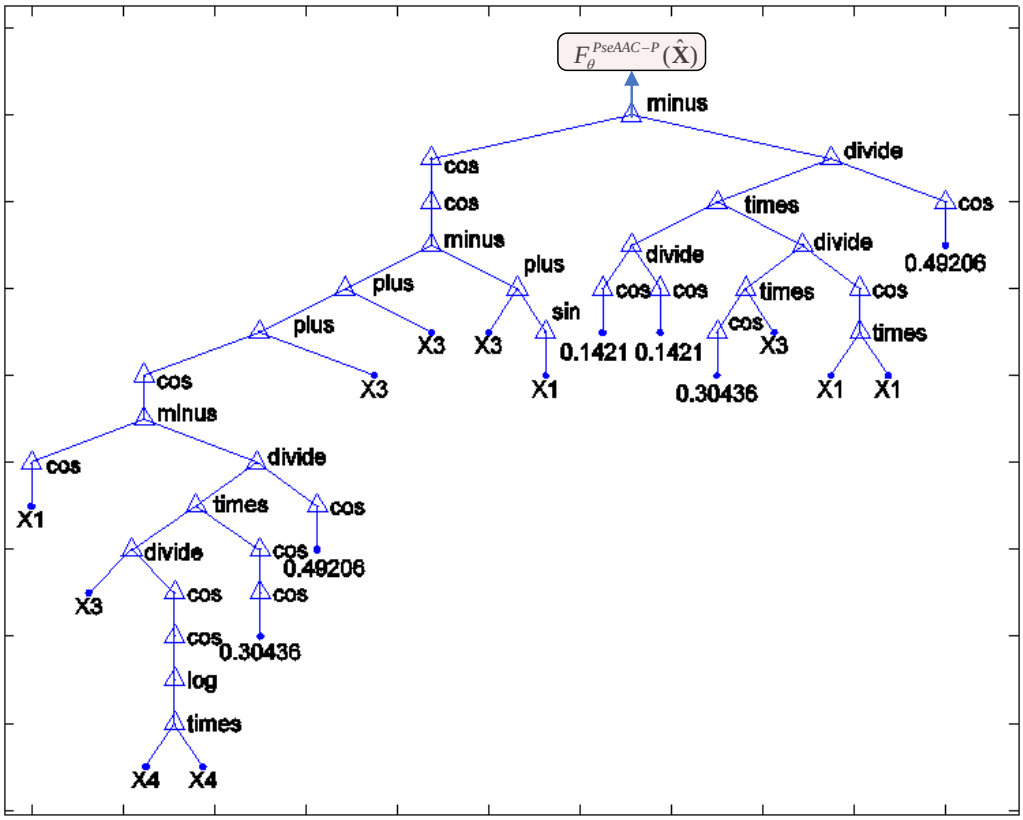


Figure 7.6 For breast cancer, tree structure of the best individual of Evo-Ens in PseAAC-P feature space.

7.3 Results and Discussion

In this section, the experiment results of the proposed Can-Evo-Ens system for C/NC and B/NBC datasets is presented. The performance of individual and proposed predictors is assessed using different quality measures in various feature spaces. The

overall results are obtained using 10-fold cross-validation data sampling technique. The overall performance of the Can-Evo-Ens system is compared with individual (base-level predictors) and approaches from previous studies.

7.3.1 Performance of Individual Predictors

Table 7.2 highlights the performance achieved, using 10-fold cross-validation data sampling technique, by individual base predictors in different feature spaces for C/NC and B/NBC datasets. For C/NC dataset, NB predictor has obtained Acc value of 93.12% in PseAAC-P and the overall highest MCC value of 74.97% in SAAC space compared to other feature spaces. KNN predictor has attained the best values of Acc 95.43%, Sn 96.53%, Sp 94.34%, G_{mean} 95.43%, F_{Score} 95.38%, and MCC 64.36% for PseAAC-S feature space. Prediction Acc 96.99% of SVM is better in AAC space compared to other spaces. However, it is observed that RF predictor in PseAAC-S

Table 7.2 Performance of individual base predictors using different feature extraction strategies.

Model/		C/NC dataset					B/NBC dataset				
	Feature Space	Acc	Sp	Sn	G_{mean}	F_{score}	Acc	Sp	Sn	G_{mean}	F_{score}
NB	AAC	90.75	82.08	99.42	90.34	91.49	93.84	88.97	98.72	93.72	94.13
	SAAC	88.96	77.92	100.0	88.27	90.05	93.47	86.94	100.0	93.24	93.87
	PseAAC-S	90.81	96.42	85.20	90.64	90.26	88.06	90.47	85.65	88.03	87.77
	PseAAC-P	93.12	86.24	100.0	92.87	93.56	94.33	90.15	98.50	94.23	94.55
KNN	AAC	93.12	93.3	92.95	93.12	93.11	92.61	93.58	91.65	92.61	92.54
	SAAC	91.04	90.75	91.33	91.04	91.07	91.86	90.47	93.25	91.85	91.97
	PseAAC-S	95.43	96.53	94.34	95.43	95.38	94.91	94.97	94.86	94.91	94.91
	PseAAC-P	94.51	96.3	92.72	94.49	94.41	95.45	94.86	96.04	95.45	95.48
SVM	AAC	96.99	95.49	98.49	96.98	97.09	94.00	91.33	96.68	93.97	94.16
	SAAC	96.65	95.26	98.04	96.64	96.69	94.11	88.22	100.0	93.93	94.44
	PseAAC-S	92.14	84.51	99.77	91.82	92.70	94.06	90.90	97.22	94.00	94.24
	PseAAC-P	96.19	93.18	99.19	96.14	96.30	94.33	89.83	98.82	94.22	94.57
RF	AAC	96.76	96.76	96.76	96.76	96.76	94.47	97.32	95.61	96.46	96.44
	SAAC	96.82	97.23	96.41	96.82	96.81	94.41	96.79	96.04	96.41	96.40
	PseAAC-S	97.92	98.26	97.57	97.91	97.91	95.34	97.64	96.04	96.83	96.82
	PseAAC-P	96.94	97.69	96.18	96.93	96.91	95.13	96.90	97.32	97.10	97.12

space has achieved the highest values of Acc 97.92%, Sn 98.26%, Sp 97.57%, G_{mean} 97.91%, F_{Score} 97.91%, and MCC 68.01%. For B/NBC dataset, KNN predictor in PseAAC-P has given the best Acc value 95.45%. SVM predictor in PseAAC-P space has provided the best Acc value 94.33%. NB predictor has gained Acc value 94.33%, again, in PseAAC-P feature space compared to other spaces.

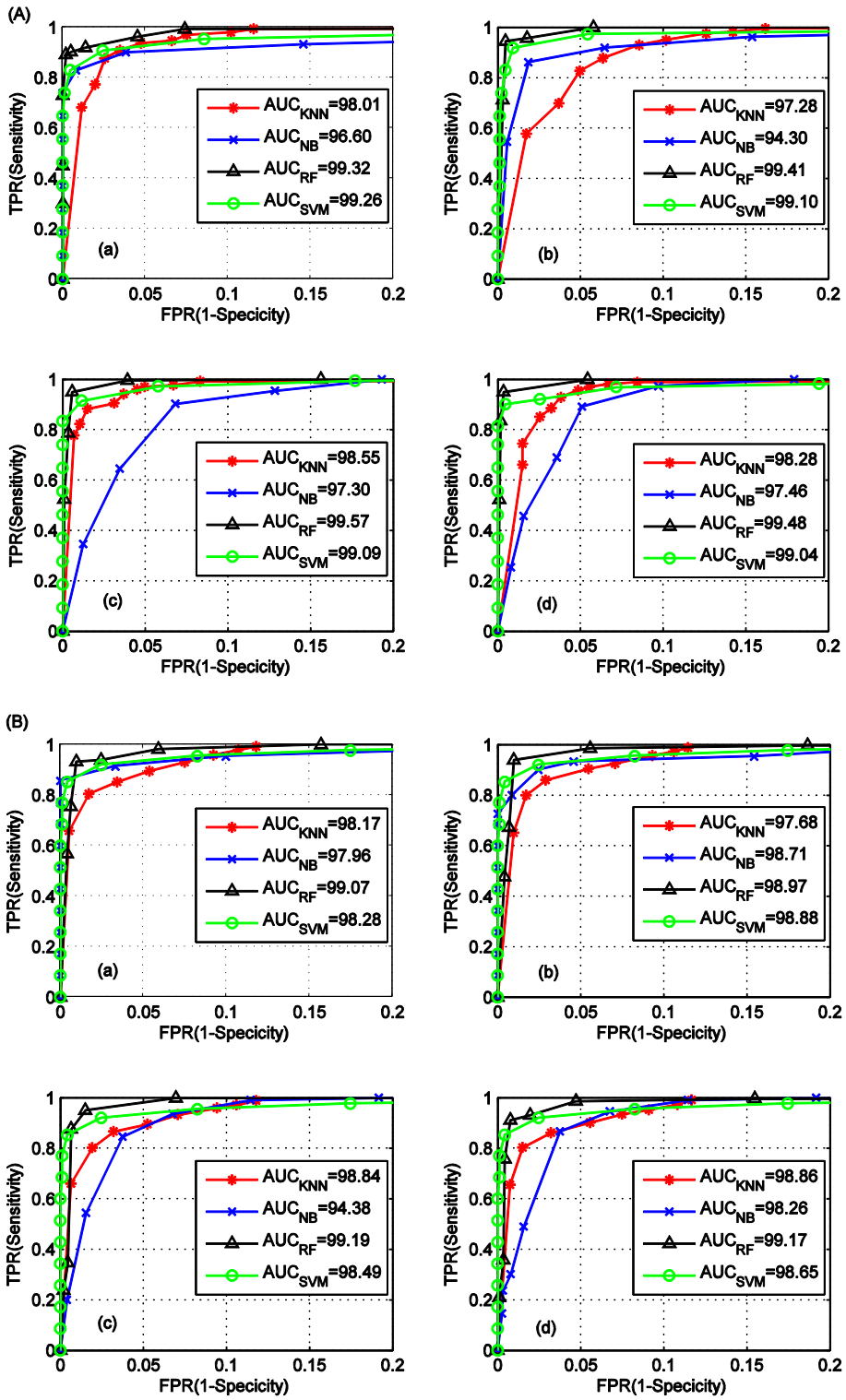


Figure 7.7 ROC curves (partial) of individual predictors, NB, KNN, SVM, and RF for (A) C/NC and (B) B/NBC datasets using: (a) AAC, (b) SAAC, (c) PseAAC-S, and (d) PseAAC-P spaces. ***(Partial ROC curves are plotted for better visualization of region of interest. High sensitivity levels are desirable in a medical decision.)***

Fig. 7.7A(a-d) and Fig. 7.7B(a-d) demonstrates the partial ROC curves of individual predictors, using 10-fold cross-validation technique, for different feature

spaces for C/NC and B/NBC datasets. From Fig. 7.7A(a-d), it is observed that PseAAC-S based RF predictor has provided the best AUC value 99.57%, followed by PseAAC-P (99.48%), SACC (99.41%), and AAC (99.32%) spaces. It is observed that predictors have provided the highest values of AUC measure in different feature spaces, for instance, NB (97.47%) in PseAAC-P space, KNN (98.55%) in PseAAC-S space, SVM (99.26%) in SAAC space. The average prediction performances of RF are higher 0.81%, 2.87%, and 2.01% than SVM, KNN, and NB predictors, respectively.

Fig. 7.7B(a-d) shows that RF has the best value of AUC for all feature spaces. From this figure, again, it is observed that PseAAC-S space has provided the best AUC value 99.19% for RF predictor. However, other predictors have provided the highest values of AUC in different spaces, for example, NB (98.71%) in SAAC space, KNN (98.86%) in PseAAC-P space, SVM (98.88%) in SAAC space.

Table 7.3 shows the values of average Q statistics of individual predictors in different feature spaces for C/NC and B/NBC datasets (2nd and 3rd columns). The lower value of Q gives the higher improvement in the proposed predictor.

Table 7.3 The values of average Q and optimal γ_{pso}^{FS} of individual predictors in different spaces.

Feature Space	Average Q		γ_{pso}^{FS}	
	C/NC dataset	B/NBC dataset	C/NC dataset	B/NBC dataset
AAC	0.3979	0.3985	0.4034	0.5018
SAAC	0.3971	0.3921	0.7593	0.5659
PseAAC-S	0.3965	0.3923	0.4141	0.6391
PseAAC-P	0.3987	0.3989	0.1383	0.7193

For the development of ensemble system, accurate and diverse characteristics of base predictors are vital. From Table 7.2 and Fig. 7.7, it is found that the performance of individual base predictors varies in different feature spaces. RF predictor in PseAAC-S space has achieved the highest Acc compared to other spaces. This accurate response of predictors is helpful for our proposed system. Additionally, the lower values of average Q-statistic (Table 7.3) highlight the generation of useful diversity of base predictors in different feature spaces. Such diverse-type of response

of individual predictors are beneficial for development of the proposed system. Through the GP evolution process, we exploit effectively the useful diversity of base predictors. As a result, the performance of the proposed system is ameliorated in different feature spaces.

7.3.2 Performance of the Proposed Evolutionary Ensemble System

The best predictions $\hat{g}_{Ens}^{FS}$ of the proposed Evo-Ens are computed with the best threshold value γ_{pso}^{FS} (Table 7.3, last two columns) for C/NC and B/NBC datasets. Table 7.4 demonstrates various performance measures of the proposed Evo-Ens in different feature spaces. For C/NC dataset, it is observed that PseAAC-S based predictor ($\hat{g}_{Ens}^{PseAAC-S}$) has achieved the highest values of Acc 99.02%, Sp 98.73% G_{mean} 99.01%, and F_{Score} 99.02%. In case of B/NBC, again, predictor $\hat{g}_{Ens}^{PseAAC-S}$ has gained the highest values of Acc 98.39%, Sp 99.36% G_{mean} 98.39%, and F_{Score} 98.38%.

Table 7.4 Performance of the proposed Evo-Ens in different feature spaces for C/NC and B/NBC datasets.

Proposed Predictor	C/NC dataset					B/NBC dataset				
	Acc	Sp	Sn	G_{mean}	F_{score}	Acc	Sp	Sn	G_{mean}	F_{score}
g_{Ens}^{AAC}	98.84	98.61	99.08	98.84	98.85	97.81	98.72	96.90	97.80	97.79
g_{Ens}^{SAAC}	98.90	98.61	99.19	98.90	98.90	97.70	97.64	97.75	97.69	97.70
$g_{Ens}^{PseAAC-S}$	99.02	98.73	99.30	99.01	99.02	98.39	99.36	97.43	98.39	98.38
$g_{Ens}^{PseAAC-P}$	98.38	97.34	99.42	98.37	98.40	98.29	98.18	98.40	98.29	98.29

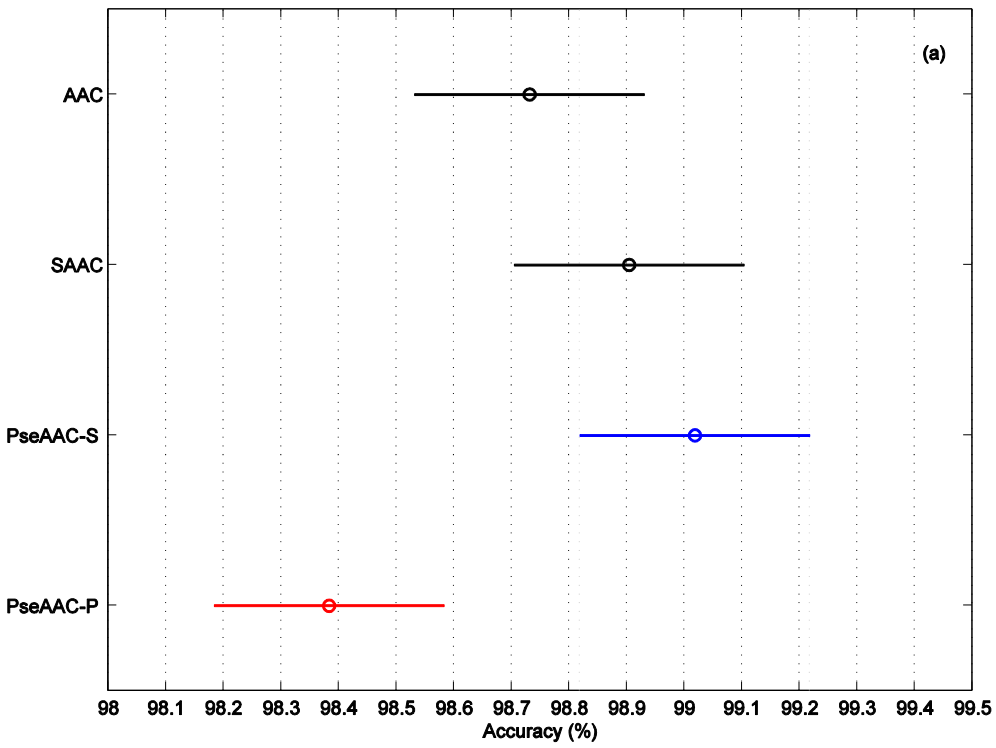
The statistical analysis is carried out using Analysis of Variance (ANOVA) to estimate the significance improvement of models in different feature spaces. Table 7.5 shows the results of this analysis for C/NC and B/NBC datasets. These results highlight a significant difference among the models because the returned p-value (0.001) for the C/NC dataset is lower than the defined α-value (0.05). Further, multiple comparison tests is performed to check whether each pair of four models is significantly different or not. The results obtained using multiple comparison procedures are given in Fig. 7.8. The graph of Fig. 7.8a indicates that PseAAC-S feature space is only significantly different from PseAAC-P. However, PseAAC-P is significantly different from AAC and SAAC spaces. In case of B/NBC dataset, the p-value is near to zero. This is evidence that the Acc performance varies from one model to another. The graph of Fig. 7.8b shows that PseAAC-S feature space is

Table 7.5 Analysis of variance (α=0.05) for the average accuracy.

Dataset	Source	Sum of squares	Degree of freedom	Mean square	F-score	*P*-value
C/NC	Models	2.30266	3	0.76755	7.22	0.001
	Acc (rows)	0.67621	9	0.07513	0.71	0.6973
	Error	2.86889	27	0.10626		
	Total	5.84776	39			
B/NBC	Models	3.58647	3	1.19549	40.14	0.0000
	Acc (rows)	0.27105	9	0.03012	1.01	0.4554
	Error	0.80406	27	0.02978		
	Total	4.66158	39			

significantly different from AAC and SAAC feature spaces for B/NBC dataset.

Additionally, to assess the scalability of the approach, we have computed the computational time for different feature spaces. The average training time of the proposed approach is computed to be 1235.8 and 570.4 sec in the feature spaces of PseAAC-S (60 dimensions) and PseAAC-P (40 dimensions), respectively, while keeping all other parameters constant. Therefore, ensemble models in PseAAC-S space consumed about twice more time than PseAAC-P space.



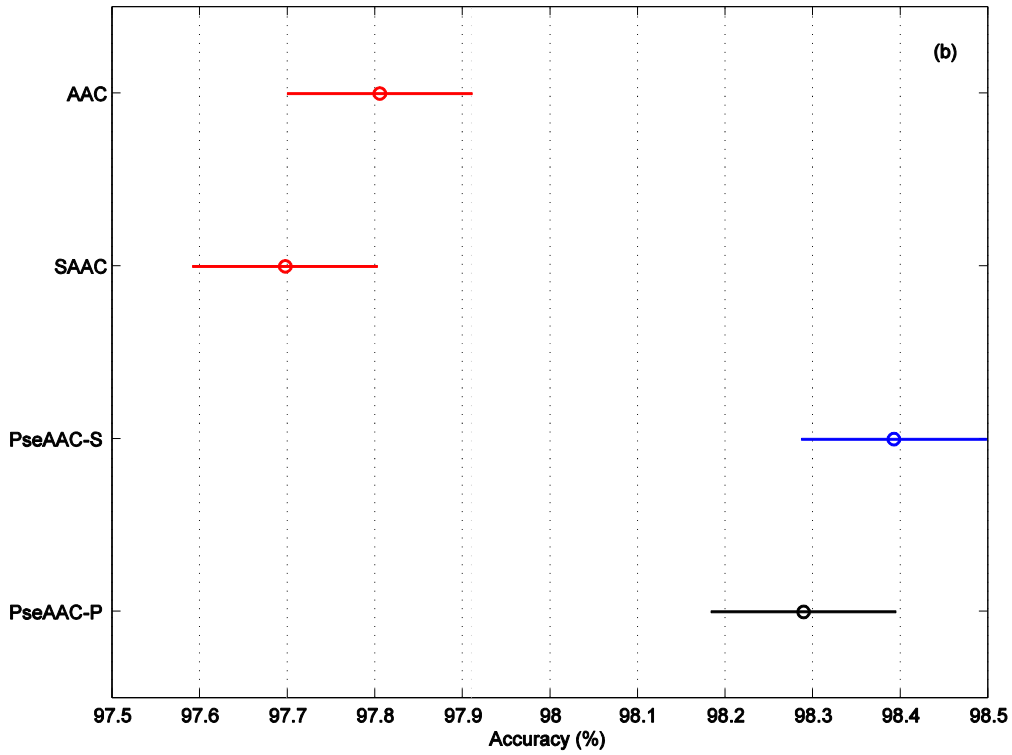


Figure 7.8 Difference in mean values of accuracy of different models using multiple comparison procedures for (a) C/NC dataset and (b) B/NBC dataset.

Improved results are observed for the proposed predictor (Table 7.4) over individual predictors (Table 7.2) using C/NC and B/NBC datasets. It is observed (Table 7.4) that using variation of amino acid compounds in cancerous protein primary sequences, the proposed predictor has attained the highest Sp values of 99.42% and 98.40% for C/NC and B/NBC datasets, respectively. For B/NBC dataset, the average Sp of the proposed approach is higher with respect to Sn measure. Hence, our predictor has predicted cancerous protein sequences more precisely. It is evident from Tables 7.2 and 7.4 that after combining the predictions of base predictors, the value of Sp and Sn considerably ameliorate. It is found (Table 7.4) that although the *Sn* decreases slightly (Table 7.2) when applying GP module, the *Sp* is improved. The higher values of Sn and Sp are desired for medical decision. G_{mean} and F_{Score} are also improved because these values depend on Sn and Sp measures.

The prediction performance of the Evo-Ens is analyzed in terms of ROC curves in different feature spaces for C/NC and B/NBC datasets. From Fig. 7.9a, for C/NC prediction, it is observed that the proposed system in PseAAC-S space has provided the best AUC value 99.95%, followed by SACC (99.93%), AAC (99.91%), and PseAAC-P (99.87%) spaces. For B/NBC prediction, Fig. 7.9b demonstrates that,

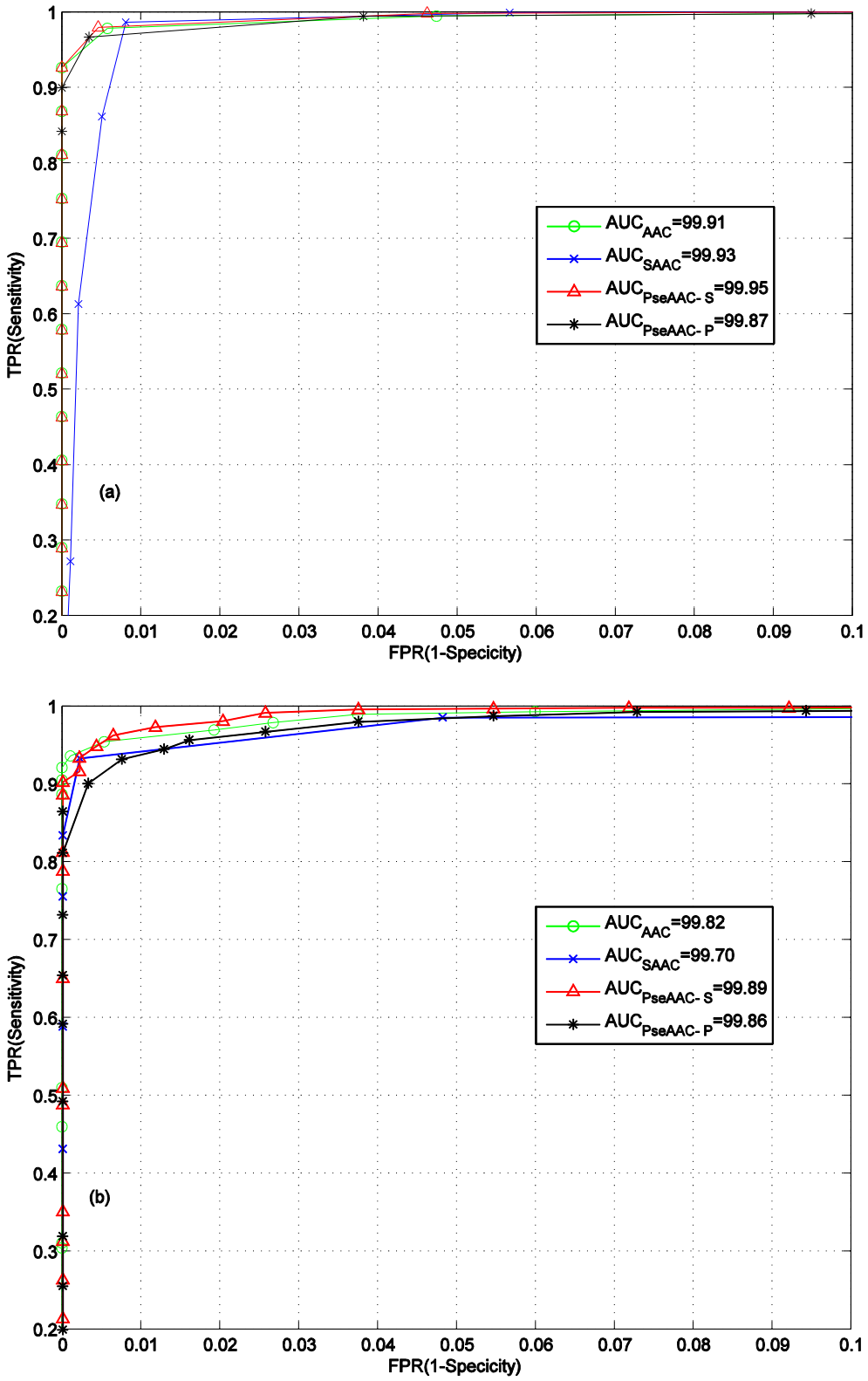


Figure 7.9 ROC curves (partial) of the proposed predictors for (a) C/NC dataset and (b) B/NBC dataset using AAC, SAAC, PseAAC-S, and PseAAC-P spaces.

again, PseAAC-S based predictor has provided the best AUC value 99.89%, followed by PseAAC-P (99.86%), ACC (99.82%), and SAAC (99.70%). Thus, Fig. 7.9 shows

the improved ROC curves of the proposed system in different feature spaces. The improved ROC curve is helpful in selecting operating point of the predictor.

Fig. 7.10 shows performance comparison of the Evo-Ens in different feature spaces for C/NC and B/NBC datasets. This figure highlighted that Evo-Ens is better for C/NC dataset in all feature spaces, except PseAAC-P. In case of PseAAC-P, the proposed system has provided similar results for both datasets.

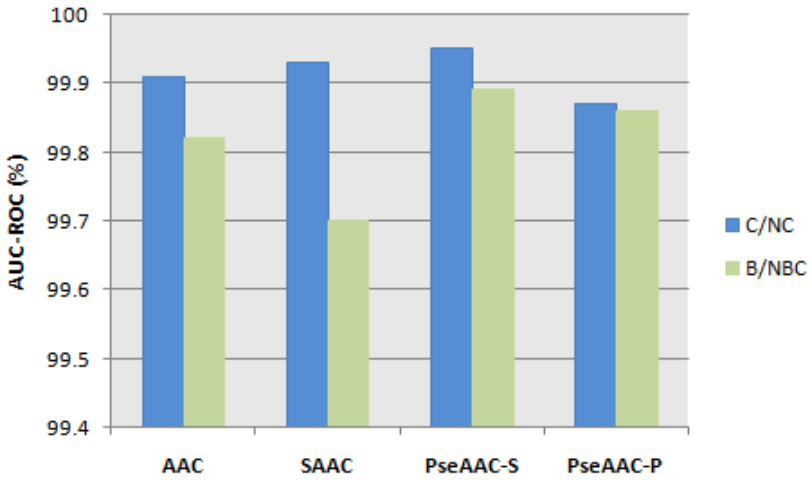


Figure 7.10 Performance comparison of the proposed system in different feature spaces.

7.3.3 Overall Performance Comparison

In this subsection, a performance comparison of the proposed approach is carried out with individual and approaches from previous studies. Tables 7.2 and 7.4 summarize the overall performance of individual predictors and the proposed predictors. Our approach outperformed individual predictors in terms of Acc, Sp, G_{mean}, and F_{Score} for C/NC and B/NBC datasets. NB predictor has shown slight progress over our predictor in terms of Sn and MCC for all feature spaces except PseAAC-S. SVM models show improvement in terms of MCC using PseAAC-S space for C/NC and B/NBC datasets. By comparing Figs. 7.7 and 7.9, it is also observed that among individual predictors, RF has achieved better AUC value 99.57% (Fig. 7.7A(c)), although our predictor (Fig. 7.9a) has attained the highest AUC value 99.95% in PseAAC-S space for C/NC dataset.

For comparison purpose, four well-known conventional ensemble approaches of AdaBoostM1, Bagging, GentleBoost, and Random Subspace are developed. AdaBoostM1 and GentleBoost are implemented using Decision Tree as base

classifiers. However, Bagging and Random Subspace are implemented using Discriminant Analysis as base learning algorithm. Fig. 7.11 shows performance comparison of the proposed ensemble system with conventional ensemble approaches in the best feature space PseAAC-S. For C/NC dataset, from Fig. 7.11a, it is evident that the proposed system has outperformed the conventional ensemble by providing the best AUC value of 99.95%. For B/NBC dataset, Fig. 7.11b highlights the comparison of the proposed approach with other ensemble approaches in the best performing feature space (PseAAC-S). Again, it is observed that our approach has outperformed the previous ensemble approaches by producing the best AUC value of 99.89%. Now, the RIA in the proposed approach is discussed.

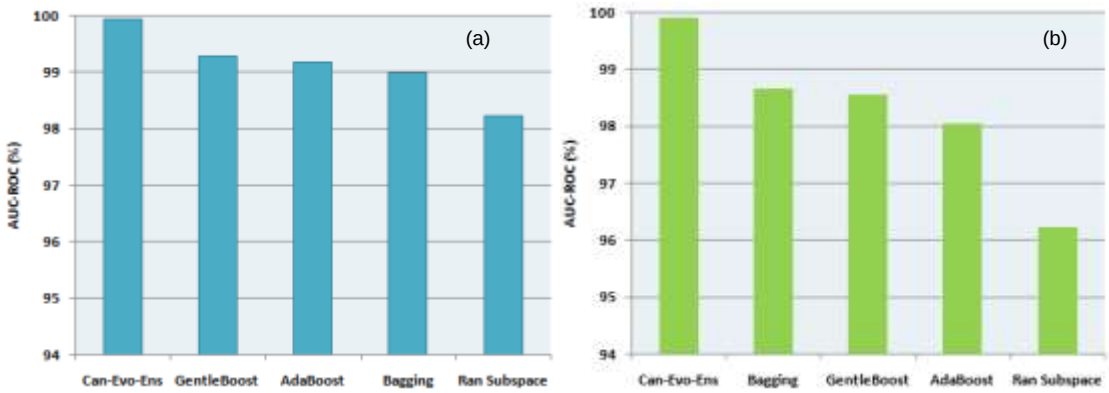


Figure 7.11 Performance comparison of the proposed ensemble system with well-known ensemble approaches in the best PseAAC-S space for (a) C/NC dataset and (b) B/NBC datasets.

Table 7.6 highlights relative improvement in results of our approach over base predictors and conventional ensemble approaches. For C/NC dataset, Table 7.6, our approach has gained the RIA in AUC-ROC measure of 10.79 % with NB, 8.10% with KNN, 4.38% with SVM, and 1.98% with RF. In terms of Acc, the RIA of 34.77 % is observed with NB, 22.63% with KNN, 13.98% with SVM, and 6.90% with RF. Our approach has attained the highest RIA over NB in Sp (61.96%), Sn (14.81%), G_{mean} (36.61 %), and F_{Score} (32.75%). In terms of MCC measure, the highest RIA of 46.72% is observed over KNN. However, the proposed approach has obtained the smallest RIA over RF predictor for AUC-ROC (1.98%), Acc (6.90%), Sp (3.45%), G_{mean} (6.91 %), and F_{Score} (6.99%) measures. For B/NBC datasets, similar performance trend is observed.

In case of conventional ensembles (Table 7.6), for C/NC dataset, the proposed

approach has obtained RIA in AUC-ROC measure of 10.01 % (AdaBoostM1), 9.62% (Bagging), 8.88% (GentleBoost), and 19.05% (Random Subspace). The RIA of accuracy values of 21.77%, 18.76%, 17.40%, and 45.82% is observed with AdaBoostM1, Bagging, GentleBoost, and Random Subspace, respectively. The proposed approach has gained the highest RIA over Random Subspace in terms of Sp (45.32%), Sn (55.38%), G_{mean} (48.05%), and F_{Score} (47.49%), and MCC (88.26%). Our approach has attained the smallest RIA over GentleBoost using AUC-ROC (8.88%), Acc (17.40%), Sp (21.53%), G_{mean} (17.43%), F_{Score} (17.11%) and over Bagging using Sn (13.40%) and MCC (6.87%). Similar behavior is observed for B/NBC datasets. On the other hand, our approach has shown sufficient improvement over previous approaches (see Table 7.7) for both of C/NC and B/NBC datasets.

Table 7.6 RIA of the proposed evolutionary approach.

	Method	RIA (C/NC Dataset) (%)					RIA (B/NBC Dataset) (%)				
		Acc	Sp	Sn	G_{mean}	F_{score}	Acc	Sp	Sn	G_{mean}	F_{score}
Base predictor	NB	34.77	61.96	14.81	36.61	32.75	24.68	42.00	9.56	25.20	24.01
	KNN	22.63	17.71	27.68	22.63	22.81	18.61	21.54	15.72	18.60	18.50
	SVM	13.98	28.07	1.53	14.41	13.10	16.67	37.37	-2.23	17.07	15.63
	RF	6.90	3.45	10.42	6.91	6.99	5.54	5.40	5.69	5.54	5.56
Conventional	AdaBoostM1	21.77	26.54	17.25	21.83	21.42	28.11	27.32	29.09	28.17	28.38
	Bagging	18.76	31.07	7.24	18.97	18.03	15.38	31.96	-0.05	15.62	14.58
	GentleBoost	17.40	21.53	13.40	17.43	17.11	20.81	28.81	13.06	20.84	20.37
	Random Subspace	45.82	45.32	55.38	48.05	47.49	36.41	21.80	56.38	37.82	39.41
Proposed approach		-	-	-	-	-	-	-	-	-	-

The conventional ensemble approaches such as AdaBoostM1, Bagging, GentleBoost, and Random Subspace have shown poor performance compared to the proposed approach due to the following.

1) The conventional ensemble approaches are developed by generating a set of classifiers that are trained from a single learning algorithm and iteratively retraining the base classifier using a subset of most informative training data.
2) The conventional ensemble approaches do not exploit effectively the useful diversity of base predictors.
3) The conventional ensemble approaches do not effectively combine the

decisions of the base predictors using protein sequence features. However, in the proposed approach, we have used diverse learning algorithms to generate diverse decision spaces and then efficiently integrate these decisions with evolutionary algorithm of GP.

4) Another limitation of the previous approaches is to have merely employed one level by taking the original input data to give a single output prediction.

In Table 7.7, accuracy comparison of the proposed approach is carried out with previous well-known approaches for breast cancer. The purpose of the comparison is to analyze/visualize which approach is better, using different classifiers and data/features, to predict more accurately breast cancer problem.This comparison could be supportive to cancer researchers for the choice of data/features/classifiers for future research of breast cancer. Optimized-LVQ and Big-LVQ models have provided accuracy near to 96.80% [113]. The prediction performance using clinical features has enhanced accuracy in the range of 97.07-97.51% for Fuzzy-GA, AR+NN, SVM+EAs, and SBS-BPPSO approaches. Ensemble (NF, KNN, QC) using information gain based selected clinical features has provided accuracy of 97.14% [117]. TIs based QPDR models have achieved maximum accuracy of 90.81 % [12]. On the other hand, for C/NC dataset, our system using PseAAC-S feature space has given the best prediction of 99.02%. Similarly, for B/NBC dataset, we have achieved the best prediction of 98.39%.

During GP evolution process, complex structure of predictor functions is developed. These functions in the form of equations and figures accentuated the functional dependency on the predictions of base predictors. Therefore, the GP has potential to exploit the most informative feature spaces, which are extracted from the numerical descriptors based on physicochemical properties of amino acids. In this study, it is found that GP based predictor in PseAAC-S space has provided the best performance for cancer prediction. This is because PseAAC-S feature space carries the most discriminant information.

The overall performance of the proposed system is superior due to two reasons. First, the utilization of most informative features derived from the physicochemical properties of amino acids. At feature level, these features have a potential to accommodate the variation of amino acid composition in cancer and

breast-cancer protein sequences with reference to non-cancer proteins. Second, our GP evolutionary approach is designed differently from conventional approaches. At decision level, our approach has optimal combined the predictions of diverse types of learning algorithms and thereby ameliorate the performance. On the other hand, the proposed study has some limitations due to the stochastic nature of the GP technique. In order to find the best parametric values from large search space, during GP training process, we had to run GP simulation several times. Further, the candidate solutions may converge slowly near the global optima.

Table 7.7 Prediction comparison of the proposed Evo-Ens with other approaches.

Approach	Feature extraction strategy		Acc (B/NBC) (%)		Dataset
Ensemble (AdaboostM1+SVM-SMO) [121]	Clinical features			82.00	UCI repository: Mammographic Mass Dataset
Boosting [66]				95.43	UCI repository: Wisconsin-breast-cancer
Bagging [66]				96.96	-Do-
Optimized-LVQ [62]				96.70	-Do-
Big-LVQ [62]				96.80	-Do-
Fuzzy-GA [116]				97.36	-Do-
AR +NN [53]				97.40	-Do-
SVM+EAs [63]				97.07	-Do-
SBS-BPPSO [122]	Entropy based clinical features			97.51	-Do-
Fuzzy-SVM [114]	Clinical features extracted with Principal Component Analysis			96.35	-Do-
Ensemble (NF KNN QC) [117]	Information gain based clinical features			97.14	-Do-
	C/NC	B/NBC	C/NC	B/NBC	
QPDR [12]	*pTIe* (embedded)	*TIe+ dTIe* (embedded)	90.00	91.80	Same as present study
RF	PseAAC-S	PseAAC-S	97.92	95.34	**Present study**
GentleBoost	AAC	PseAAC-S	96.36	93.74	-Do-
Bagging	PseAAC-S	PseAAC-P	94.54	94.75	-Do-
Proposed approach	PseAAC-S	PseAAC-S	99.02	98.39	-Do-

In this chapter, classifier stacking based evolutionary ensemble system for reliable prediction of breast cancer is discussed. This study revealed that PseAAC-S feature space has yielded excellent discrimination power over other feature extraction strategies. Comparative analysis has demonstrated that evolutionary approach has performed superior than conventional ensemble approaches of AdaBoostM1, Bagging, GentleBoost, and Random Subspace. The results have shown that if all the individual predictors are optimized even then the evolutionary approach is capable of improving the performance.

Chapter 8: Conclusions and Future Work

Development of an accurate and reliable automated cancer prediction system can provide adequate information for cancer diagnosis and drug discovery. Development of intelligent decision making ensemble systems requires protein amino acid sequences to be represented as numerical descriptors through the use of informative and discriminative feature extraction strategies. In this thesis, various IDME approaches based on discriminative feature extraction strategies are developed. Further, IDME systems have exploited the feature spaces generated by protein molecular descriptors to enhance the overall prediction performance. The proposed approaches have attained significant improvement compared to state-of-the-art approaches reported in the literature.

8.1 Conclusions

In the following paragraphs, conclusive remarks of Chapter 3-7 are presented.

Chapter 3 focused on the development of various individual prediction systems for proteins linked with human breast and colon cancers. Together with the development of individual systems, MTD technique to handle imbalanced data is discussed. Performance of the models is analyzed in different feature spaces of AAC, SAAC, PseAAC-S, and PseAAC-P. From the comparison, it is inferred that MTD-SVM is better than KNN, NB, and QPDR models in terms of prediction performance. The MTD-KNN model established as a benchmark that has achieved moderately better prediction as compared to MTD-NB. The proposed $SVM_{PseAAC\text{-}S}$ model has given the best discriminant feature space of PseAAC-S. The $SVM_{PseAAC\text{-}S}$ prediction model yielded accuracy of 96.71% for C/NC dataset, 95.18% for B/NBC dataset, and 96.50% for CC/NCC dataset. In terms of AUC/AUCH measures, for PseAAC-P feature space, SVM models are found best for the prediction of cancer disease. However, using PseAAC-S features, SVM models are most suitable for B/NBC and CC/NCC. The proposed approach has indicated that MTD-KNN and MTD-SVM based individual systems can be used as a tool for the prediction of cancer.

Chapter 4 explored the notion of oversampling and cost learning of various homogeneous ensemble systems to handle the imbalanced data. The analysis has demonstrated that random ensemble system CanPro-IDSS achieved maximum values of AUC of 99.79% for C/NC dataset and 99.58% for B/NBC dataset. Overall, results highlighted that the proposed system using balanced data outperformed CSL based system. This system is excellent because: (i) it uses the most informative feature spaces generated from physicochemical properties of amino acids, (ii) it incorporates MTD as preprocessor for data balancing, and (iii) RF employs random data sub-sampling and ensemble strategies. A web predictor "CanPro-IDSS" is developed.

Chapter 5 dedicated for the development of heterogeneous ensemble system by exploiting decision spaces of different classifiers. The main consideration was to develop improved intelligent decision making system IDMS-HBC for cancer prediction. Improved performance is achieved by integration of diverse predictions of several classifiers through majority voting strategy. The proposed system achieved AUC of 99.79% for C/NC and 99.58% for B/NBC datasets. From analysis, it is observed that the IDMS-HBC system is superior to the individual, conventional ensembles, and other state-of-the-art approaches.

Chapter 6 focused on the exploitation of discriminative power of different feature spaces, which are extracted from protein primary sequences using physicochemical properties of Hd and Hb of amino acids. Further, the variation of amino acid molecules associated with protein cancer is studied. It is observed that Proline, Serine, Tyrosine, Cysteine, Arginine, and Asparagine amino acids offer adequate discrimination for cancerous and healthy proteins. Analysis demonstrated that RF, SVM, and KNN based ensemble are more effective than their individual counterparts in different feature spaces. The proposed IDM-PhyChm-Ens has achieved the best accuracy of 99.48% and 97.63% for ensemble-RF and ensemble-SVM, respectively. It is observed that combined feature spaces of SAAC+PseAAC-S and AAC+SAAC show the best discrimination using ensemble-RF and ensemble-NB. The comparative analysis highlights the improved performance of the proposed IDM-PhyChm-Ens system over existing approaches.

Chapter 7 provided a deeper insight about stacking based evolutionary system. It is explored that the performance of the Can-Evo-Ens depends on useful information

extracted from protein primary sequences in different feature spaces. The proposed Can-Evo-Ens system has demonstrated its robustness for independent validation dataset. The proposed system using PseAAC-S space has achieved the highest values of accuracies 99.02% and 98.39% for C/NC and B/NBC datasets, respectively. This approach has yielded excellent discrimination power in PseAAC-S space over other feature extraction strategies. In PseAAC-S space, the proposed system has provided the highest AUC values of 99.95% and 99.89% for C/NC and B/NBC datasets, respectively. Comparative analysis has demonstrated that this approach has performed better than conventional ensemble approaches of AdaBoostM1, Bagging, GentleBoost, and Random Subspace. It is inferred that even if all the individual predictors are optimized, the proposed predictor is capable of improving the overall performance.

Further, one aim of this study was to determine which classifier has more discriminant exploitation of certain types of feature spaces, that is to find effective and the best combination of "*classifier+feature-space(s)*" for cancer prediction. In following paragraphs, an analysis of the results of the core chapters (3-7) along with reasons to why a particular combination of base classifiers with a particular space of features performing best in a given setting is discussed.

For protein sequences, different feature spaces exhibited different characteristics and yielded different prediction results. So, instead of using conventional feature selection and transform based extraction strategies, we used sequence-order and correlation based feature generation strategies. Protein sequence data contain intrinsic dependencies between their constituent elements. Given a protein sequence over the amino acid alphabets, the dependencies between neighboring elements are modeled by generating all the contiguous present in sequence data.

Usually, the proposed approaches presented in this thesis have potential to exploit the most informative feature spaces, which are extracted from the numerical descriptors based on physicochemical properties of amino acids. All amino acids have different physicochemical properties owing to their differences in side chains. Features derived from physicochemical properties of protein sequences are quite helpful in cancer prediction. In this study, it is found that the proposed predictors in

PseAAC feature spaces, particularly PseAAC-S space, have provided the best performance over other feature spaces for cancer prediction. This is because, in other feature spaces such as AAC, important position-specific or hidden information in protein sequences are lost. On the other hand, PseAAC feature space reflects better sequence-order information and length of a protein sequence for prediction. The PseAAC feature space has capability to reflect the inherent correlation with corresponding target labels and carries the most discriminant information. This discriminant information is due to effective use of Hd and Hb properties of amino acids of Proline, Tyrosine, Serine, Arginine, Asparagine, Isoleucine, and Cysteine (see Fig. 6.1). Thus at feature level, this feature space has a potential to accommodate the variation of amino acid composition in cancer and breast-cancer protein sequences with reference to non-cancer proteins.

In chapters 4 and 7, a statistical analysis is carried out using Analysis of Variance (ANOVA) to estimate the significance improvement of the proposed models in different feature spaces. Tables 4.6 and 7.5 show the results of this analysis. It is inferred from this analysis that PseAAC-S feature space is emerging as the best space. It is also observed from other core chapters that almost in all proposed system, PseAAC-S feature space with the combination of our proposed algorithms has shown the best performance. It is concluded that PseAAC-S feature space has the most discriminant information than other feature spaces and hence it is independent of our proposed algorithms.

Our proposed algorithms performed better with the combination of PseAAC-S feature space because:

1) Typically, proposed approaches are designed differently from conventional approaches.

2) The proposed algorithms have good generalization capability on balanced datasets, which are balanced by MTD technique. These approaches generated effective search space and decision space for ameliorated prediction of cancer.

3) The proposed algorithms have used more effective and different types of combining strategies to integrate the decisions of the base predictors.

4) The proposed algorithms have effectively generated and exploited the useful diversity of base predictors. As a result, the performance of the proposed approaches is ameliorated using PseAAC-S feature space.

Overall, the proposed IDME systems could easily be utilized by clinicians for the diagnosis of cancer using protein sequences of the affected tissue. The protein sequences of amino acids can be obtained from DNA sequences, mass spectroscopy, Edman degradation, etc. These protein sequences could simply be fed to IDME systems. If the protein is related to cancer i.e., pattern of amino acids in protein is changed, it will diagnose cancer patient, otherwise non-cancer patient (see publication No.3). Once the cancer is diagnosed by our system, the clinician may proceed further by recommending other standard tests to know the severity of the cancer.

8.2 Future Work

The proposed IDME systems have achieved significant improvement over existing state-of-the-arts approaches. The highest performance of the classifier is reported for PseAAC based feature spaces using $20+i\times\lambda$ discrete components, with λ=20 and i=2. Infuture, increased the number of discrete components ($\lambda > 20$) can be explored for enhnced discrimination that might be useful for accurate cancer prediction. Further in future, to enhance the performance of the proposed ensemble approaches, filter or wrapper based features selection techniques can be incorporated. During training phase, useful feature spaces or classification models can be combined through weighted majority scheme.

In the current thesis, physicochemical properties of amino acids in protein primary sequences of Hd and Hb are used to construct different feature spaces. In future, more physicochemical properties of amino acids can be employed. Additionally, other pre-processing methods prior to the feature generation strategies could be explored in order to study their effectiveness in discriminating cancer protein molecules.

Prediction performance of IDME systems depends on the discrimination power of feature spaces generated from protein amino acid sequences. Generally, the prediction system depends on the successful development of ensemble model. Therefore, it is suggested that the development of a new or modified feature

extraction strategy and ensemble model might enable the predictor to achive improved performance. Additionally, performance measures such as “equal error rates” along with AUC may be utilized.

In future, instead of using oversampling technique of MTD in minority class, the undersampling of majority class can be employed using clustering or bootstrap random sampling technique and then computing the mean of the clustered samples. This process is repeated until the size of the minority-class equals the size of majority-class samples and thereby producing balanced the data.

In this thesis, the proposed IDME systems were developed for four protein feature spaces of AAC, SAAC, PseAAC-S, and PseAAC-P. However, in future work, more protein sequences and new feature spaces could be employed.

The research has contributed towards an enhanced understanding of the relationship between protein sequences and various cancers. Consequently, proteomic analysis may provide adequate information for diagnosis, prevention, and therapy of the cancer. One primary aspect of this work has involved analyzing the role of variation of amino acid molecules in cancer related protein using physicochemical properties. This methodology could be easily extended to other types of cancer diseases. It is expected that this study would also be helpful in the future research of sequential information such as any nucleic acid sequence, translational-bioinformatics, medical informatics, genome, proteome, drug discovery, etc. An additional benefit may be to develop automated systems that are able to extract features from online datasets of cancerous protein to facilitate research for specific cancer diseases.

References

[1] D. Caroline, K. Brasseur, V. Leblanc, S. Parent, É. Asselin, and G. Bérubé, "SAR study of tyrosine–chlorambucil hybrid regioisomers; synthesis and biological evaluation against breast cancer cell lines," *Amino acids,* vol. 43, pp. 923-935, 2012.

[2] J. Gallagher, "James G: Predicted global cancer cases," in *Reference WHO GloboCan*. vol. 2014 London, 2014.

[3] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray, "Cancer Incidence and Mortality Worldwide," International Agency for Research on Cancer, Lyon, France, Available from http://globocan.iarc.fr. 2013.

[4] B. M. Good, S. Loguercio, O. L. Griffith, M. Nanis, C. Wu, and A. I. Su, "The Cure: Making a game of gene selection for breast cancer survival prediction," *arXiv preprint arXiv,* vol. 1402, p. 3632, 2014.

[5] R. Alteri, C. Barnes, and et al, "American Cancer Society ", Atlanta, Georgia 2014.

[6] J. Milenković, K. Hertl, A. Košir, J. Žibert, and J. F. Tasič, "Characterization of spatiotemporal changes for the classification of dynamic contrast-enhanced magnetic-resonance breast lesions," *Artificial Intelligence in Medicine,* vol. 58, pp. 101-114, 2013.

[7] O. L. Griffith, P. François, M. E. Oana, M. H. Laura, A. C. Eric, T. S. Paul, and W. G. Joe, "A robust prognostic signature for hormone-positive node-negative breast cancer," *Genome medicine,* vol. 5, 2013.

[8] A. A. Margolin, E. Bilal, E. Huang, T. C. Norman, L. Ottestad, B. H. Mecham, B. Sauerwine, and et al, "Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer," *Science translational medicine,* vol. 5, pp. 181re1-181re1, 2013.

[9] C. Pierrick, A. P. Joseph, P. Poulain, A. G. d. Brevern, and J. Rebehmed, "Cis-trans isomerization of omega dihedrals in proteins," *Amino acids,* vol. 45, pp. 279-289, 2013.

[10] Y. Ji-Yeon, K. Yoshihara, K. Tanaka, M. Hatae, H. Masuzaki, H. Itamochi, M. Takano, K. Ushijima, J. L. Tanyi, G. Coukos, Y. Lu, G. B. Mills, and R. G. W. Verhaak, "Predicting time to ovarian carcinoma recurrence using protein markers," *The Journal of Clinical Investigation,* vol. 123, pp. 3740–3750, 2013.

[11] R. G. Ramani and S. G. Jacob, "Improved Classification of Lung Cancer Tumors Based on Structural and Physicochemical Properties of Proteins Using Data Mining Models," *PLoS ONE,* vol. 8, p. e58772, 2013.

[12] C. R. Munteanu, A. L. Magalhães, E. Uriarte, and H. González-Díaz, "Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices," *Journal of theoretical biology,* vol. 257, pp. 303-311, 2009.

[13] K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics,* vol. 21, pp. 10-19, 2005.

[14] D. Cahill, K. Kinzler, B. Vogelstein, and C. Lengauer, "Genetic instability and darwinian selection in tumours," *Trends Cell Biol.,* vol. 9, pp. M57-60, 1999.

[15] J. Marx, "Debate surges over the origins of genomic defects in cancer," *Science,* vol. 297, pp. 544-546, 2002.

[16] S. Naron, H. Lynch, T. Conway, P. Watson, J. Feunteun, and G. Lenoir, "Increasing incidence of breast cancer in family with BRCA1 mutation," *Lancet.,* vol. 341, pp. 1101-1102, 1993.

[17] E. Hulleman and K. Helin, "Molecular mechanisms in gliomagenesis," *Adv. Cancer Res,* vol. 94, pp. 1-27, 2005.

[18] S. Irena, J. Livsey, J. A. Keanec, and G. Nenadic, "Text mining of cancer-related information: Review of current status and future directions," *International journal of medical informatics,* vol. 83, pp. 605-623, 2014.

[19] A. Khan, A. Majid, and C. Tae-Sun, "Predicting protein subcellular location: exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers," *Amino Acids,* vol. 38, pp. 347-350, 2010.

[20] M. Hayat and A. Khan, "Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition," *Journal of Theoretical Biology,* vol. 271, pp. 10-7, 2011.

[21] J. A. Benediktsson and P. H. Swain, "Consensus theoretic classification methods," *IEEE Transactions on Systems, Man Cabernet,* vol. 22, p. 688−704, 1992.

[22] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?," *Machine Learning,* vol. 54, p. 255−273, 2004.

[23] D. Wolpert, "Stacked generalization," *Neural Networks,* vol. 5, pp. 241-259, 1992.

[24] A. Khan, A. Majid, and M. Hayat, "CE-PLoc: An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition.," *Computational Biology and Chemistry,* vol. 35, pp. 218-229, 2011.

[25] A. Safdar, A. Majid, and A. Khan, "IDM-PhyChm-Ens: Intelligent decision-making ensemble methodology for classification of human breast cancer using physicochemical properties of amino acids," *Amino acids,* vol. 46, pp. 977-993, 2014.

[26] A. Majid, A. Safdar, I. Mubashar, and K. Nabeela, "Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines," *Computer Methods and Programs in Biomedicine,* vol. 113, pp. 792-808, 2014.

[27] L. Breiman, "Bagging predictors," *Machine Learning,* vol. 24, pp. 123-140, 1996.

[28] R. Schapire, "The strength of weak learnability," *Machine Learning,* vol. 5, pp. 197-227, 1990.

[29] L. Shapley and B. Grofman, "Optimizing group judgmental accuracy in the presence of interdependencies, Public Choice," vol. 43, pp. 329-333, 1984.

[30] C. K. Chow, "Statistical independence and threshold functions, IEEE Transactions on Electronic Computers EC-," vol. 14, pp. 66-68, 1965.

[31] B. V. Dasarathy and B. V. Sheela, "A composite classifier system design: conceptsand methodology," in *Proceedings of the IEEE,* 1979, pp. 708–713.

[32] L. Rastrigin and R. H. Erenstein, *Method of Collective Recognition*. Energoizdat, Moscow, 1981.

[33] L. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 12, pp. 993-1001, 1990.

[34] L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 22, pp. 418-435, 1992.

[35] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognition,* vol. 29, pp. 341-348, 1996.

[36] T. Ho, J. J. Hull, and S. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 16, pp. 66-75, 1994.

[37] Y. Freund, "Boosting a weak learning algorithm by majority," *Information Computing,* vol. 121, pp. 256-285, 1995.

[38] M. Kearns and U. Vazirani, *An Introduction to Computational Learning Theory*. MA, USA: MIT Press, Cambridge, 1994.

[39] D. Angluin, "Queries and concept learning," *Machine Learning,* vol. 2, pp. 319-342, 1988.

[40] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *Journal of Artificial Intelligence Research,* vol. 10, pp. 271-289, 1999.

[41] F. J. Ordóñez, A. Ledezma, and A. Sanchis, "Genetic Approach for Optimizing Ensembles of Classifiers," FLAIRS Conference, 2008.

[42] Y. Chen, W. Man-Leung, and L. Haibing, "Applying Ant Colony Optimization to configuring stacking ensembles for data mining," *Expert Systems with Applications,* vol. 41, pp. 2688-2702, 2014.

[43] Y.-H. Huang, Y.-C. Chang, C.-S. Huang, T.-J. Wu, J.-H. Chen, and R.-F. Chang, "Computer-aided diagnosis of mass-like lesion in breast MRI: Differential analysis of the 3-D morphology between benign and malignant tumors," *Computer Methods and Programs in Biomedicine,* vol. 112, pp. 508-517, 2013.

[44] D. James and A. T. Balaban, *Topological indices and related descriptors in QSAR and QSPAR*: CRC Press Llc, 2000.

[45] G. Ferino, H. González-Díaz, G. Delogu, G. Podda, and E. Uriarte, "Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative proteome-disease relationships (QPDRs) and predicting prostate cancer," *Biochemical and Biophysical Research Communications* vol. 372, pp. 320-5, 2008.

[46] W.-Y. Cheng, T.-H. O. Yang, and D. Anastassiou, "Development of a prognostic model for breast cancer survival in an open challenge environment," *Science translational medicine,* vol. 5, pp. 181ra50-181ra50, 2013.

[47] M. Emmanuel, M. M. Alvarez, and V. Trevino, "Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm," *Computational Biology and Chemistry,* vol. 34, pp. 244-250, 2010.

[48] R. Saima, M. Hussain, A. Ali, and A. Khana, "A Recent Survey on Colon Cancer Detection Techniques," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 10, pp. 545 - 563, 2013.

[49] J. Dheeba, N. A. Singh, and S. T. Selvi, "Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach," *Journal of biomedical informatics,* vol. In press, 2014.

[50] F. Gorunescu and S. Belciug, "Evolutionary strategy to develop learning-based decision systems. Application to breast cancer and liver fibrosis stadialization," *Journal of biomedical informatics*, vol. In press, 2014.

[51] G. Santanu, A. Mukherjee, S. Sengupta, and P. K. Dutta, "Cancer classification from gene expression data by NPPC ensemble," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, pp. 659-671, 2011.

[52] A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A. R. Razavi, and L. G. Ahmad, "Using three machine learning techniques for predicting breast cancer recurrence," *Journal of Health & Medical Informatics*, vol. 4, pp. 124-, 2013.

[53] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Systems with Applications*, vol. 36, pp. 3465-3469, 2009.

[54] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996.

[55] D. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data," *Artificial Intelligence in Medicine*, vol. 16, pp. 149-169, 1999.

[56] A. Kalinli, F. Sarikoc, H. Akgun, and F. Ozturk, "Performance comparison of machine learning methods for prognosis of hormone receptor status in breast cancer tissue samples," *Computer Methods and Programs in Biomedicine*, vol. 110, pp. 298-307, 2013.

[57] R. Liao, T. Wan, and Z. Qin, "Classification of benign and malignant breast tumors in ultrasound images based on multiple sonographic and textural features," in *Proceedings International Conference on Intelligent Human-Machine Systems and Cybernetics 2011 (IHMSC-2011)*, Hangzhou, 2010, pp. 71-74.

[58] F. Aminzadeh, B. Shadgar, and A. Osareh, "A robust model for gene analysis and classification," *The International Journal of Multimedia & Its Applications*, vol. 3, pp. 11-20, 2011.

[59] M. Xin, J. Guo, H. Liu, J. Xie, and X. Sun, "Sequence-Based Prediction of DNA-Binding Residues in Proteins with Conservation and Correlation Information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, pp. 1766-1775, 2012.

[60] F. K. Ahmad, S. Deris, and N. H. Othman, "The inference of breast cancer metastasis through gene regulatory networks," *Journal of biomedical informatics*, vol. 45, pp. 350-362, 2012.

[61] C. Binghuang and X. Jiang, "A novel artificial neural network method for biomedical prediction based on matrix pseudo-inversion," *Journal of Biomedical Informatics*, vol. 48, pp. 114-121, 2014.

[62] D. E. Goodman, L. Boggess, and A. Watkins, "Artificial immune system classification of multiple-class problems," in *In Proceedings of the artificial neural networks in engineering* 2002, pp. 179–183.

[63] S. Ruxandra and C. Stoean, "Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection," *Expert Systems with Applications*, vol. 40, pp. 2677-2686, 2013.

[64] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, pp. 25-41, 2000.

[65] M. Ebrahimi, E. Ebrahimie, and N. Shamabadi, "Are there any differences between features of proteins expressed in malignant and benign breast cancers?," *J Res Med Sci,* vol. 15, pp. 299-309, 2010.

[66] D. Lavanya and K. U. Rani, "Ensemble Decision Making System for Breast Cancer Data," *International Journal of Computer Applications,* vol. 51, pp. 0975 -8887, 2012

[67] Y. Liu, W. Cheng, and Z. Lu, "Decision tree based predictive models for breast cancer survivability on imbalance data," in *In Proceedings of International Conference on Bioinformatics and Biomedical Engineering. Beijing: IEEE,* Beijing, 2009, pp. 1-4.

[68] K. J. Wang, B. Makond, and K. M. Wang, "An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data," *BMC Medical Informatics and Decision Making* vol. 13, p. 124, 2013.

[69] W. Zhang, F. Zeng, X. Wu, X. Zhang, and R. Jian, "A comparative study of ensemble learning approaches in the classification of breast cancer metastasis," in *Bioinformatics, Systems Biology and Intelligent Computing, IJCBS'09, International Joint Conference on.*, 2009, pp. 242-245.

[70] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artif Intell Med,* vol. 34, pp. 113-127, 2005.

[71] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest.," *BMC medical informatics and decision making* vol. 11, p. 51, 2011.

[72] P. D. Dobson, Y. D. Cai, B. J. Stapley, and A. J. Doig, "Prediction of protein function in the absence of significant sequence similarity," *Current Medicinal Chemistry,* vol. 11, pp. 2135-42, 2004.

[73] T. Sjoblom, S. Jones, L. D. Wood, W. Parsons, J. Lin, T. D. Barber, and et al., "The consensus coding sequences of human breast and colorectal cancers," *Science,* vol. 314, pp. 268-74, 2006.

[74] C. Chen, X. Zhou, Y. Tian, X. Zou, and P. Cai, "Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network," *Analytical Biochemistry,* vol. 357, pp. 116-21, 2006.

[75] C. Tanford, "Contribution of hydrophobic interactions to the stability of the globular conformation of proteins," *Journal of the American Chemical Society* vol. 84, pp. 4240-4247, 1962.

[76] T. P. Hopp and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences," *National Acad Sciences,* vol. 78, pp. 3824-3828, 1981.

[77] V. Vapnik, *The nature of statistical learning theory*: Springer, 1999.

[78] D. Dumitru, "Prediction of recurrent events in breast cancer using the Naive Bayesian classification," *Annals of the University of Craiova-Mathematics and Computer Science Series,* vol. 36, pp. 92-96, 2009.

[79] R. Nisbet, J. Elder, and G. Miner, *Handbook of statistical analysis and data mining applications*. Burlington, MA: Academic Press, 2009.

[80] D. F. Specht, "Probabilistic neural networks," *Neural Networks,* vol. 1, pp. 109-118, 1990.

[81] A. Balmain, J. Gray, and B. Ponder, "The genetics and genomics of cancer." vol. 33: Nature Publishing Group, 2003, pp. 238-244.

[82] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," New York: Springer, 2001.

[83] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences,* vol. 55, pp. 119-139, 1997.

[84] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics,* vol. 28, pp. 337-407, 1998.

[85] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection* vol. 1: MIT press, 1992.

[86] K. Asifullah, S. F. Tahir, A. Majid, and T.-S. Choi, "Machine learning based adaptive watermark decoding in view of anticipated attack," *Pattern Recognition,* vol. 41, pp. 2594-2610, 2008.

[87] M. M. Tariq, A. Majid, and T.-S. Choi, "Optimal depth estimation by combining focus measures using genetic programming," *Information Sciences,* vol. 181, pp. 1249-1263, 2011.

[88] A. Majid, C.-H. Lee, M. T. Mahmood, and T.-S. Choi, "Impulse noise filtering based on noise-free pixels using genetic programming," *Knowledge and information systems,* vol. 32, pp. 505-526, 2012.

[89] A. Majid, M. T. Mahmood, and T.-S. Choi, "Optimal composite depth function for 3D shape recovery of microscopic objects." Microscopy research and technique," *Microscopy research and technique,* vol. 73, pp. 657-661, 2010.

[90] F. Tom, "ROC graphs: Notes and practical considerations for researchers," *Machine Learning,* vol. 31, pp. 1-38, 2004.

[91] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-57, 2002.

[92] T. M. Ha and H. Bunke, "Off-line, handwritten numeral recognition by perturbation," *Pattern Analysis and Machine Intelligence,* vol. 19, pp. 535-9, 1997.

[93] D. C. Li, C. S. Wu, T. I.Tsai, and Y. S. Lina, "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," *Computers & Operations Research,* vol. 34, pp. 966-82, 2007.

[94] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent data analysis,* vol. 6, pp. 429-449, 2002.

[95] G. M. Weiss and F. J. Provost, "Learning when training data are costly: the effect of class distribution on tree induction," *J Artif Intell Res (JAIR),* vol. 19 pp. 315-54, 2003.

[96] S. L. Phung, A. Bouzerdoum, and G. H. Nguyen, *Learning pattern classification tasks with imbalanced data sets*: InTech, 2009.

[97] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the 14th International Conference on Machine Learning* San Francisco, CA, USA: Morgan Kaufmann, 1997.

[98] P. Domingos, "Metacost: a general method for making classifiers cost-sensitive," in *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 1999.

[99] D. C. Li, C. W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets," *Computers in Biology and Medicine,* vol. 40, pp. 509-18, 2010.

[100] K. C. Chou and C. T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology,* vol. 30, pp. 275-349, 1995.
[101] A. Khan, A. Majid, and M. Hayat, "CE-PLoc: An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition," *Computational Biology and Chemistry,* vol. 35, pp. 218-229, 2011.
[102] H. Mohabatkar, "Prediction of cyclin proteins using Chous pseudo amino acid composition," *Protein and Peptide Letters,* vol. 17, p. 1207, 2010.
[103] A. Zubair, J. M. Schuemie, J. C. V. Blijderveen, E. F. Sen, C. Miriam, Sturkenboom, and J. A. Kors, "Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records," *BMC medical informatics and decision making,* vol. 13, p. 2013.
[104] A. Endo, T. Shibata, and H. Tanaka, "Comparison of seven algorithms to predict breast cancer survival," *Int J Biomed Soft Comput Hum Sc,* vol. 13, pp. 11-16, 2008.
[105] V. D. Vijver, J. Marc, D. H. Yudong, J. v. t. V. Laura, D. Hongyue, A. H. Augustinus, W. D. Voskuil, and J. S. George, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine,* vol. 347, pp. 1999-2009, 2002.
[106] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, and D. Talantov, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet,* vol. 365, pp. 671-679, 2005.
[107] S. Hanash, "Disease proteomics," *Nature* vol. 422, pp. 226-232, 2003.
[108] H. Hondermarck, C. Tastet, I. E. Yazidi-Belkoura, R.-A. Toillon, and X. L. Bourhis, "Proteomics of breast cancer: the quest for markers and therapeutic targets," *Journal of proteome research,* vol. 7, pp. 1403-1411, 2008.
[109] R. H. Alvarez, V. Valero, and G. N. Hortobagy, "Emerging targeted therapies for breast cancer," *Journal of Clinical Oncology,* vol. JCO-2009, 2010.
[110] J. M. Phang and W. Liu, "Proline metabolism and cancer," *Frontiers in bioscience: a journal and virtual library,* p. 1835, 2012.
[111] A. Richardson, "Proline Metabolism in Metastatic Breast Cancer," in *Proline Metabolism in Metastatic Breast Cancer*, 2011.
[112] M. M. R. Krishnan, S. Banerjee, C. Chakraborty, and A. K. Ray, "Statistical analysis of mammographic features and its classification using support vector machine," *Expert Systems with Applications,* vol. 37, pp. 470-478, 2010.
[113] B. Ster and A. Dobnikar, "Neural networks in medical diagnosis: Comparison with other methods," in *proceedings of the international conference on engineering applications of neural networks*, 1996, pp. 427–430.
[114] D. C. Li, C. W. Liu, and S. C. Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets," *Artificial Intelligence in Medicine,* vol. 52, pp. 45-52, 2011.
[115] K. P. Bennett and J. A. Blue, "A support vector machine approach to decision trees," in *Neural Networks Proceedings. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, 1998, pp. 2396-2401.
[116] C. A. Pena-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," *Artificial Intelligence in Medicine,* vol. 17, pp. 131-155, 1999.

[117] H. Sheau-Ling, S.-H. Hsieh, P.-H. Cheng, C.-H. Chen, K.-P. Hsu, I.-S. Lee, Z. Wang, and F. Lai, "Design Ensemble Machine Learning Model for Breast Cancer Diagnosis," *Journal of medical systems,* vol. 36, pp. 2841-2847, 2012.

[118] L. Davis, "Adapting operator probabilities in genetic algorithms," in *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, 1989, pp. 61-69.

[119] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networkw*, 1995, pp. 1942–1948.

[120] A. Majid, "Optimization and combination of classifiers using Genetic Programming," in *Faculty of Computer Science*. vol. PhD Pakistan: GIK institute, 2005.

[121] S.-T. Luo and B.-W. Cheng, "Diagnosing breast masses in digital mammography using feature selection and ensemble methods," *Journal of medical systems,* vol. 36, pp. 569-577, 2012.

[122] M.-L. Huang, Y.-H. Hung, and W.-Y.Chen, "Neural Network Classifier with Entropy Based Feature Selection on Breast Cancer Diagnosis," *J Med Syst,* vol. 34, pp. 865-873, 2010.