IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 February 2023, accepted 11 March 2023, date of publication 20 March 2023, date of current version 24 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3259907

RESEARCH ARTICLE

IDriveGenes: Cancer Driver Genes Prediction Using Machine Learning

YASIR ALI[1], MUHAMMAD SARDARAZ[2], MUHAMMAD TAHIR[2], HELA ELMANNAI[3], MONIA HAMDI[3], AND AMEL KSIBI[4]

[1]Department of Computer Science, Sir Syed Case Institute of Technology, Islamabad 4600, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan
[3]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia
[4]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia

Corresponding author: Muhammad Tahir (m_tahir@cuiatk.edu.pk)

This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R125), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT The development of high throughput sequencing technologies i.e. Next Generation Sequencing (NGS) is revolutionizing the exploration of cancer. Though sequence datasets are highly complex, mutation can occur randomly in DNA or RNA sequences that can make cells sicker or less fit. The unusual growth and behavior of genes in cells cause cancer. Cancer-driver gene cells grow when mutation occurs. Identification of cancer driver genes is a critical and challenging issue for researchers. In the proposed work, initially, robust features are extracted from the sequence dataset through Position Relative Incidence Matrix (PRIM) integrated with Accumulative Absolute Position Incidence Vector (AAPIV) generation. PRIM and AAPIV convert the single-dimensional sequence data into 2-dimensional numeric data. Support Vector Machine (SVM), Neural Network (NN), and Random Forest (RF) are used to train the model. The proposed model is validated with different validation methods i.e., independent testing, k-fold cross-validation, self-consistency, and jackknife testing. The proposed model predicts whether the given primary structure corresponds to cancer driver genes or not. Results analyses show 95%, 92%, and 69% accuracy on RF, Artificial Neural Networks (ANN), and SVM respectively. The comparative analysis with existing state-of-the-art models i.e., 20/20+ and Multimodal Deep Neural Network by integrating Multi-dimensional Data (NDNNMD) shows that the proposed model outperforms the existing techniques.

INDEX TERMS Accuracy, cancer, driver genes, machine learning, NGS.

I. INTRODUCTION

The massive parallel and high-throughput sequencing platform known as Next Generation Sequencing (NGS) technologies enforce progressive demand on statistical models and bioinformatics applications to manage and analyze intensive data produced [1]. Bioinformatics domain includes sequence analysis, gene annotation, gene expression, protein analysis, protein structure prediction, high sequence image analysis, mutation analysis in cancer, etc [2]. Mutation in genomics data usually causes cancer. Mutation can occur by Single Nucleotide Variants (SNVs), structural variants, and insertion or deletion of genes. The mutation is the change in the DNA sequence that occurs in the bases of the sequence. Mutation occurs due to DNA replication and environmental factors such as smoking, radiation, and sunlight also affect the DNA sequences. This change occurs in protein sequences that can be bad or good for health. The mutation that occurs in inheritance usually has a positive effect. However, mutation disturbs normal genes and causes diseases like cancer. A somatic cell is any cell of a living body that rises after conception. Somatic mutation is variation in any cell other than reproductive cells such as gamete, gametocyte, or germ cell [3].

Oncogenes (OG) are genes that help to grow mutated cells. When the mutation occurs in any cell, OG is activated and

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

starts growing. A large number of copies become bad genes when activated. When this happens, the cell grows rapidly and can progress to cancer disease. The Tumor Suppressor Genes (TSG) or anti-oncogenes are common genes that prevent the division of cells and repair DNA errors [4]. When the mutation occurs in a cell, it activates and prevents the growth of cancer mutation. When TSG is deactivated or does not work properly, the mutated cell can grow extensively, which can cause cancer. Driver genes containing driver mutations can be discovered from cancer mutation facts with or without earlier knowledge of pathways or further information on genetics and or protein interactions [5], [6]. This technique works when driver and passenger mutations observe the same frequencies. Additionally, it is found correct that sub-network can recognize small recurrence cancer driver genes.

NGS technologies produce a massive amount of genetic data that needs powerful computational devices, high-capacity memory, and specified software and hardware to address the particular problem i.e., driver genes prediction [7]. Cancer is a disease that has multipart interaction among environmental and genetic factors that organize carcinogenesis [8] affecting millions of people around the world. The research study presented in [9] shows that about 8.9 million deaths have been counted in 2015 and are expected to increase this number to 14.6 million by 2035. Cancer disease is coordinated by disturbance of regular cellular function. The mutation that affects the cell genetically causes the pathways deregulation that controls the fundamental process of cells [10]. The initial stage of cancer is triggered by the pile of several genetic mutations that are major causes of signaling pathways deregulation which affects cell growth, DNA restoration, and apoptosis [11]. As stated, the pathways have been deregulated, cancer cells start to grow without any normal restriction. Therefore, such a technique needs to be formed that identifies cancer-muted genes efficiently that could help to find the structure of genes involved in cancer growth. The techniques used for this purpose must be accurate along with the other parameter. The need for more accurate and efficient techniques motivated the development of the proposed model. In this article, we worked on the identification and prediction of driver genes whose mutations cause tumors or cancer. This identification could help us find a structure of cancer that is valuable for the development of novel drugs. The remaining paper is organized as follows. Section II covers related works including a detailed discussion of methods used, datasets, etc. Materials and methods are presented in section III including subsections for each component of the proposed model. Prediction algorithms with necessary details are presented in section IV. Section V presents the results and discussion. Validation methods covering details of each method are presented in section VI. Comparative analysis of the prediction models is presented in section VII covering the detailed results of each model. Comparison of the proposed model with existing methods is presented in section VIII. Section IX presents conclusion and future research directions.

II. LITERATURE REVIEW

Detection of cancer driver gene mutation in NGS data gained the attention of researchers. NGS is a massively parallel sequence technique that generates a huge amount of data that needs efficient models and frameworks for managing and analyzing the intensive data. Machine learning is used in many applications for prediction and detection including cancer driver genes [12]. There exist review articles covering the details of different methods with the pros and cons of the available methods [13], [14], [15]. This section presents methods more related to the proposed model. Research article [7] presents a MapReduce paradigm for processing parallel NGS data by distributing a repository of the DNA segment created on similar features. In research, article [16] the authors presented a dimensionality reduction approach using feature selection taxonomy of labeled, unlabeled and partially labeled gene expression microarray data for better prediction, scalability, understandability, and fitness simplification of the classifier. Feature selection in gene expression data supports cancer classification with better performance. Dimensionality reduction help to reduce storage and computational complexity, but feature selection has complex stages that are generally expensive [17]. Cancer is a heterogeneous and complex disease and various factors of environmental and genetic nature contribute to the cause of the disease. With the development of sequencing technology, a huge data on cancer genomics has been produced through various platforms such as NGS, cancer genome atlas [18], Cancer Cell Line Encyclopedia (CCLE) [19], and International Cancer Genome Consortium (ICGC), etc. These sequencing data allow the researchers to understand the mechanism of molecules and the pathogenesis basis of cancer [5]. A major challenge is the detection and distinction of driver genes that are the factor of cancer development. The earliest attempts are to identify individual driver genes that have recurring mutations [20]. Still, such methods cannot be considered for complex mutation heterogeneity in cancer genomes that have various gene mutations. Due to this a great amount of attention has been given to the assessment of mutation recurrences in genomics to create pathways that have already protein-protein interaction or known networks [21], [22]. These are habitually giving a share inside tumor cells that may cause carcinogenic possessions, e.g., metastasis, angiogenesis, or cell proliferation [23]. A key subject is that the interaction between the biological pathways and the human protein network is complete. It is of great importance to investigate innovative approaches that do not depend on prior knowledge to find pathways or mutated novel driver gene groups.

A lot of research work exists on cancer cell oncological research that indicates a low number of acquired biochemical, molecular, and cellular features. These are the main causes of alteration of key pathways that might look like a strong generalization [10]. There are 100+ distinctive types of cancer not counting additional subtypes of malignancies that have been acknowledged. But, certain directions and rules handled the transformation of human cells to cancerous

ones [24]. Cancer types need to identify specific mutations in the human genome, that can be found in many cancer types. Several researchers also attempted to classify genes that are critical to carcinogenesis into different classes based on malignant phenotypes in various experimental prototypes. There are two types of gene classes i.e., tumor suppressor and oncogenes. Cancer is caused by the activation of oncogenes as well as the inactivation of tumor suppressor genes, which later on deliberate the irregular function that specifies cancer sickness [25].

In the early stages, oncogenes called proto-oncogenes are altered by presiding mutations which gain the proliferation to a regular cell. These genetic components are known as oncogenes in their altered form and increase the proliferative ability of cells. In contrast, tumor-suppressor genes are changed and inhabited. Central cellular processes are transformed in cancer cells due to alterations in one as well as the other classes of genes e.g. metabolism, proliferation, growth, and death [7]. The mutations of these pathways give malignant and cancerous cells that can grow in huge numbers and form tumors at the local site [8]. The uncontrollable growth of these cells can occur by sidestepping the regulatory effects of the numerous mechanism which exist in a cell, controlled by key proto-oncogenes and tumor suppressor genes [9]. Homegrown cancers develop into carcinomas when they fold away and attack external tissues in the body. Authors in [26] present a Multimodal Deep Neural Network by integrating Multi-dimensional Data (MDNNMD) for diagnosing breast cancer from multidimensional data i.e., gene expression profile data and Copy Number Alteration (CNA) profile data. The small sample size or high dimensionality data may cause bad results [27]. Initially, they select effective features from gene expression profile data that include approximately 24,000 genes and CNA profile data that include approximately 26,000 genes using mRMR method [28]. The mRMR feature selection method reduces the dimensionality of data effectively without the loss of important information. The said method selects 400 genes and 200 genes from gene expression profile data and CNA profile data respectively to fit the Deep Neural Network (DNN) prediction model. DNN is applied to extract information effectively from data by greedily training each DNN on each sub-data.

Research paper [29] presents deepDriver framework that predicts cancer driver genes using a Convolutional Neural Network (CNN) based on somatic mutations. Support Vector Machine (SVM) and 20/20+ [30] methods were also applied to rank unknown genes. These methods combined with CNN improve the accuracy of the proposed approach. The experiment was made on d Lung Adenocarcinoma (LUAD), Colon Adenocarcinoma (COAD), and Breast Invasive Carcinoma (BRCA) genomes data and obtained significant scores. For further investigation said approaches can be applied to the state-of-the-art data such as pan-cancer to predict driver genes with better accuracy. Genomic Analysis of Mutations Extracted by Sequencing (GAMES) tool presented in [31] identifies and annotates mutation using NGS technology. GAMES enable the reduction of the complexity of huge DNA sequence data. This tool allows a detailed investigation of genetic and mining functional mutation of different NGS platforms. GAMES helps to extract information about divergence and make available genome annotation integrated with a genomic database. DriverML presented in [32] uses a machine learning approach that incorporates a supervised learning algorithm and weighted scoring test to identify cancer driver genes. The supervised approach scores the functional values of alteration of DNA sequences and integrates with various mutation types in somatic cells. The weighted score statistics that link all mutations can universally test each protein sequence across the genome and quantify the functional impact of various mutation forms on the protein. DriverML was applied to 31 cancer mutation datasets from TCGA and compared with 20 other common tools as the benchmark. The research article [33] presents IMaxDriver framework for driver genes prediction. It is a network-based tool using the maximization algorithm on the human transcriptional regularity network. Initially assigned weight and pruned TRN via the use of tumor-specific genes. Then find each gene's impact by influencing the maximation approach. The uppermost genes with the maximum influence rate are selected as likely driver genes. IMaxdriver identifies 408 driver genes collectively including new driver genes. Mutation can be predicted from extra-biological information about the sequence and structure of protein determined by the mutated gene. MutationAssessor presented in [34] combines protein area information with an evolutionary conservation model to identify the functional impact of somatic mutations. OncodriveFM [35] and TransFIC [36] use ML algorithms trained on known cancer mutations and focus on potential driver mutations. OncodriveFM recognizes genes by high functional mutations. A machine learning approach called 20/20+ proposed in [30] differentiates driver genes from passenger mutation in cancer. It utilizes the random forest tree trained on known cancer driver genes to recognize cohort-level cutoffs that fit the said type of identification. This method needs previous molecular knowledge for the alteration. Another approach presented in [37] presents LOTUS, a machine learning-based method to predict cancer driver genes by combining mutation frequency, functional impact, and pathway-based features. The COSMIC CGCv86 dataset is used for training and tested on the complete COSMIC database which contains 19,320 genes. Identifying new driver genes is still a major problem and challenge for the researcher. However, several approaches have been applied and lots of methods have been developed to recognize them. Number of various tools such as MutSigCV [38], 20/20+ [30], MuSic [39], TUSON [40], OncodriveFML [35], FunSeq [41], ANNOVAR [42], IntOGen [43], and CHASM [44], etc. have been proposed to identify gene mutation in sequenced data. Driver genes having driver mutations can be predicted using cancer mutation facts. This technique work when driver and

passenger mutations observe the same frequencies. More, it is found accurate that sub-network can recognize small recurrence cancer driver genes. Moreover, biological information about the protein structure and sequence encrypted by the mutated gene can predict functional mutation impact [21], [45]. These techniques are applied to the non-silent SNVs to changes in the corresponding proteins, an amino acid sequence.

III. METHODS AND MATERIAL

The details of the proposed model are presented in this section. The dataset of cancer driver genes is used to validate the proposed model. To remove the redundancies, pre-processing was used first. Features vectors were generated using the pre-processed dataset. These feature vectors were used to train various classifiers. Different validation techniques were used to test each classifier. In the proposed system, extracting a robust feature vector from sequence data is the core part to fit in the machine learning prediction model. Feature extraction means converting or transforming the input datasets into feature vector form. All the attributes and features in the datasets are mostly not used, so features that perform better roles in prediction are extracted from the dataset. The feature vector of n-dimension contains numerical values as features of an object. Several feature extraction techniques such as Position Relative Incidence Matrix (PRIM), Reverse Position Relative Incidence Matrix (RPRIM), Accumulative Absolute Position Incidence Vector (APPIV), and Reverse Accumulative Absolute Position Incidence Vector Generation (RAAPIV) have been used to extract novel features. Then statistical moments such as raw, central, and Hahn moments are applied to find further significant properties of massive data. These feature vectors are further used to train various classifiers i.e., Random Forest (RF) classifiers, Artificial Neural Network (ANN), and SVM. ANN has interconnected layers of neurons and the ANN is based on a feed-forward network and uses a back-propagation algorithm to reduce errors. Another prediction algorithm used in the research is the RF classifier for predicting cancer driver genes and non-cancer driver genes. In the end, SVM is used to predict cancer driver genes. Figure 1 presents the workflow of the proposed model IDriveGenes.

A. DATASET COLLECTION

The dataset has been obtained from NCBI, a free repository containing genetic data with biological functions [46]. Advanced search option was used to download positive samples for both cancer and non-cancer genes. Negative notation was used to search negative data samples. After collecting data from NCBI, CD-hit suite [47] is used for clustering. Analogous clusters were generated for both samples with sequence identity parameters of 60%. With the help of which 763 positive and 1805 negative cancer driver genes clusters were left. To balance the data random oversampling is used. It is a common method of oversampling in NGS nucleotide data. To balance the class distribution, this method duplicates instances of the minority class in the dataset at random. This can be useful when the minority class is under-represented in the dataset. Equation 1 shows the sum of positive and negative data generated from dataset samples.

$$S = S^{+} \cup S^{-} \tag{1}$$

where S^{+} represents the positive data sample and S^{-} shows negative data. $\cup$ represents union operation.

B. FEATURE EXTRACTION

In feature extraction, the one-dimensional input datasets are conversed and transformed into two-dimension features in vector form. Following steps are performed in sequence to achieve robust feature vectors.

1) GENE MAPPING

The feature vector contains numerical values as features and consists of n dimensions. The combination of nucleotides builds up genes and is also used in making DNA and RNA which is the genetic code of species. In gene expression, the DNA is first transcribed into mRNA. For performing the specific protein functions, the RNA may act directly or be the starting material for the synthesis. In the process of phenotypic trait inheritance, the genes are transformed into offspring. The genotype of an organism is responsible for the appearance of phenotypic along with several developmental and environmental factors. The polygenic associations between the genes and the external environment control the biological traits despite all the complexities. All the characteristics are not immediately visible like the possibility of illnesses, blood type, or thousands of essential biochemical processes that make up life while some are all immediately visible, such as skin color or several limbs. So, for the feature extraction, some mathematical model is required that considers K–tuple nucleotide composition and position in the gene sequence [48]. Gene sequences can contain characters from these 4 alphabets i.e., A = {A, G, T, C}

Gene Sequences for Dinucleotide i.e.

A = {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}

Gene Sequences for Trinucleotide i.e.

A = {AAA, AAC, AAG, AAT, ACA, ACG, ACT, ACC, AGA, AGC, AGT, AGG, ATA, ATC, ATG, ATT, CAA, CAC, CAG, CAT, CCA, CCG, CCT, CCC, CGA, CGC, CGT, CGG CTA, CTC, CTG, CTT, GAA, GAC, GAG, GAT, GCA, GCG, GCT, GCC, GGA, GGC, GGT, GGG, GTA, GTC, GTG, GTT, TAA, TAC, TAG, TAT, TCA, TCG, TCT, TCC, TGA, TGC, TGT, TGG, TTA, TTC, TTG, TTT}

The dataset contains an array of these alphabets and RF can be applied to this format because it is a good representation for input. When the input is well-formed, RF can better learn the relationships between the data to predict invisible sequences.

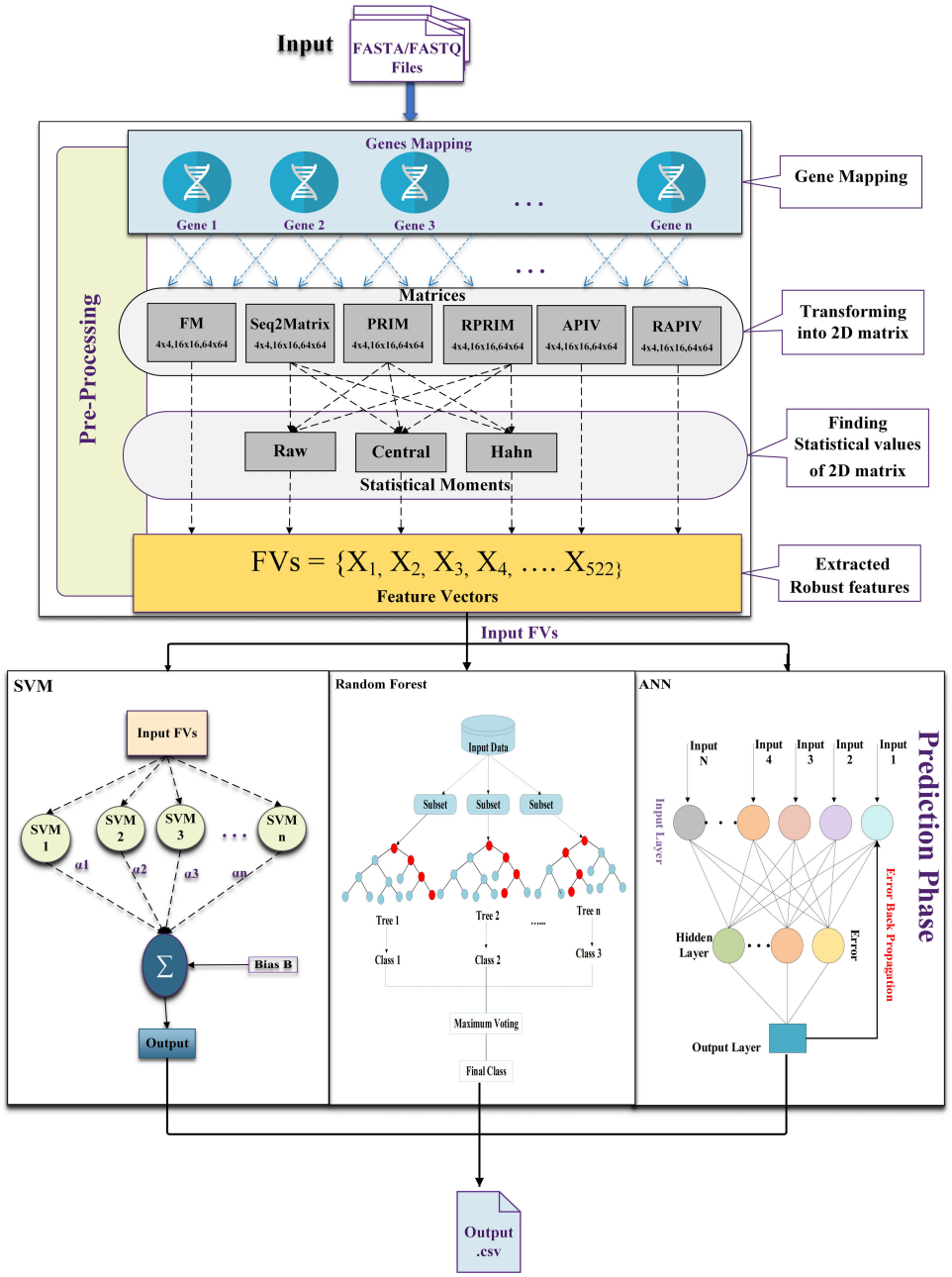

FIGURE 1. Flow of proposed model.

2) POSITION RELATIVE INCIDENCE MATRIX (PRIM)

Since the position of the amino acid in the polypeptide chain is very important, PRIM indicates the relative position of the K-tuple of the nucleotide chain [49]. Here we draw a relative probability matrix up to 64×64 positions to make the system more efficient. PRIM is a $n \times n$ matrix as shown in Equation 2.

$$S_{PRIM}n = \begin{bmatrix} s_{1\to1} & s_{1\to2} \cdots & s_{1\to j} \cdots & s_{1\to n} \\ s_{2\to1} & s_{2\to2} \cdots & s_{2\to j} \cdots & s_{2\to n} \\ s_{i\to1} & s_{i\to2} \cdots & s_{i\to j} \cdots & s_{i\to n} \\ s_{N\to1} & s_{N\to2} \cdots & s_{N\to j} \cdots & s_{N\to n} \end{bmatrix} \quad (2)$$

A component $S_{i\to j}$ has a relative position of the j^{th} element related to the occurrence of i^{th} element. If the matrix is of 4×4 dimensions, then the total is 16 coefficients, if the matrix is of 16×16 dimensions, then there are 256 coefficients in total, and if the matrix is of 64×64 dimensions, one is a total of 4096 coefficients. The capacity is huge, so we provided them with PRIM4, PRIM16, and PRIM64 as input, counting the moments to reduce them up to 24.

3) REVERSE POSITION RELATIVE INCIDENCE MATRIX (RPRIM)

RPRIM [49] extracts hidden traits from protein sequences with homologous sequence ambiguity. RPRIM has a 4×4, 16×16, and 64×64 matrices containing 16, 256, and 4096 coefficients respectively, as in PRIM, but is used for

inverse sequences. RPRIM is an $n \times n$ matrix as shown in Equation 3.

$$S_{RPRIM}n = \begin{bmatrix} s_{1\to1} & s_{1\to2}\cdots & s_{1\to j}\cdots & s_{1\to n} \\ s_{2\to1} & s_{2\to2}\cdots & s_{2\to j}\cdots & s_{2\to n} \\ s_{i\to1} & s_{i\to2}\cdots & s_{i\to j}\cdots & s_{i\to n} \\ s_{N\to1} & s_{N\to2}\cdots & s_{N\to j}\cdots & s_{N\to n} \end{bmatrix} \tag{3}$$

4) DETERMINING FREQUENCY MATRIX

A genetic model indicates the amount of time a nucleotide is present in a DNA sequence. Therefore, it is important and beneficial to detect the number of nucleotide sites. To calculate the frequency matrix denoted as ξ to represent the maximum number of K-tuples produced by nucleotides is computed using Equation 4 [49].

$$\xi = \{\tau_1, \tau_2, \cdots, \tau_i\} \tag{4}$$

The i^{th} value is 4 for nucleotide, 16 for dinucleotide, and 64 for trinucleotide. The frequency matrix helps us to know how the sequence is limited through different frequencies of each nucleotide in the series.

5) ACCUMULATIVE ABSOLUTE POSITION INCIDENCE VECTOR GENERATION (AAPIV)

A frequency matrix indicates the frequency of amino acids and tells us how the sequence is generated. It does not provide a residual relative position that can help us to find information about the nucleotide composition of genes. The aggregated frequency matrix does not provide relative position information, so an aggregated location frequency vector called Accumulative AAPIV is created to obtain the required information. AAPIV is a 4-element vector in which all numerical values of the nucleotide appear in basic order with their respective locations as given by Equation 5 [49].

$$K = \{\mu_1, \mu_2, \cdots, \mu_i\} \tag{5}$$

where the i_{th} element of AAPIV4, AAPIV16, and AAPIV64 is computed using Equation 6.

$$\mu_i = \sum_{k=1}^{n} p_k \tag{6}$$

6) REVERSE ACCUMULATIVE ABSOLUTE POSITION INCIDENCE VECTOR GENERATION (RAAPIV)

The feature extraction method extracts hidden and interesting patterns, and the AAPIV method is also used to accomplish the same task. RAAPIV is used to extract useful and hidden information about the relative position of residues in the sequence [49]. RAAPIV was developed by reversing the primary sequence of DNA and then AAPIV is generated from the reversed sequence. RAAPIV is a 4-element vector, 16 is a dinucleotide, and 64 is a trinucleotide as presented by Equation 7

$$Z = \eta_1, \eta_2, \ldots, \eta_i \tag{7}$$

where, η_i is the i_{th} element of AAPIV4, AAPIV16, AAPIV64 is computed using Equation 8.

$$\eta_i = \sum_{k=1}^{n} p_k \tag{8}$$

C. STATISTICAL MOMENTS

The statistical moment is a symbolic measure that describes the appearance of the data distribution. There are many kinds of moments, and each describes certain attributes of the data. Some moments describe the size of the data, while others describe the direction of the data. In this research, Hahn, central moment, and original moment are used to address the problem [48]. The original moments are used to estimate the mean and standard deviation of the data as they have no scaling and positional variations. Constant scale is a feature that is not affected by the scale of adding any length, energy, or other variables. The same position means that it is not affected by the movement of data values. The central moment is like the original moment as it provides the same information. They are scale-invariant but calculated along the centroid of the data. Hahn moments use polynomial values as their moment scores. They are neither position-invariant nor scale-invariant. In this research work, non-scale invariant moments are used. Since every moment has its technology, the data is presented in this way. The data is used in a two-dimensional format at every moment, so one-dimensional data is converted into a two-dimensional format. Suppose we have a genetic sequence P as presented by Equation 9.

$$P = a_1, a_2, a_3, \ldots\ldots, a_K \tag{9}$$

where, k represents series of residues, and a_i is the i_{th} K-tuple nucleotide sequence. As a result, an $n \times n$ dimensional matrix is formed to describe all amino acid components as shown in Equation 10.

$$P' = \begin{bmatrix} p_{1\to1} & p_{1\to2}\cdots & p_{1\to n} \\ p_{2\to1} & p_{2\to2}\cdots & p_{2\to n} \\ \vdots & \vdots & \vdots \\ p_{n\to1} & p_{n\to2}\cdots & p_{n\to n} \end{bmatrix} \tag{10}$$

where, P' is used for moment computation. For raw moments computation, Equation 11 is used.

$$M_{ij} = \sum_{p=1}^{n}\sum_{q=1}^{n} p^i q^j \beta_{pq} \tag{11}$$

where i and j represent the degree of the moment and β_{pq} is an arbitrary element. Moments are computed up to 3 degrees and are expressed as $M_{00}, M_{01}, M_{02}, M_{03}, M_{10}, M_{11}, M_{12}, M_{20}, M_{21}$, *and* M_{30}. The centroids are used in central moment computation as expressed in Equation 12.

$$\eta_{ij} = \sum_{p=1}^{n}\sum_{q=1}^{n} (p-\bar{x})^i (q-\bar{y})^j \beta_{pq} \tag{12}$$

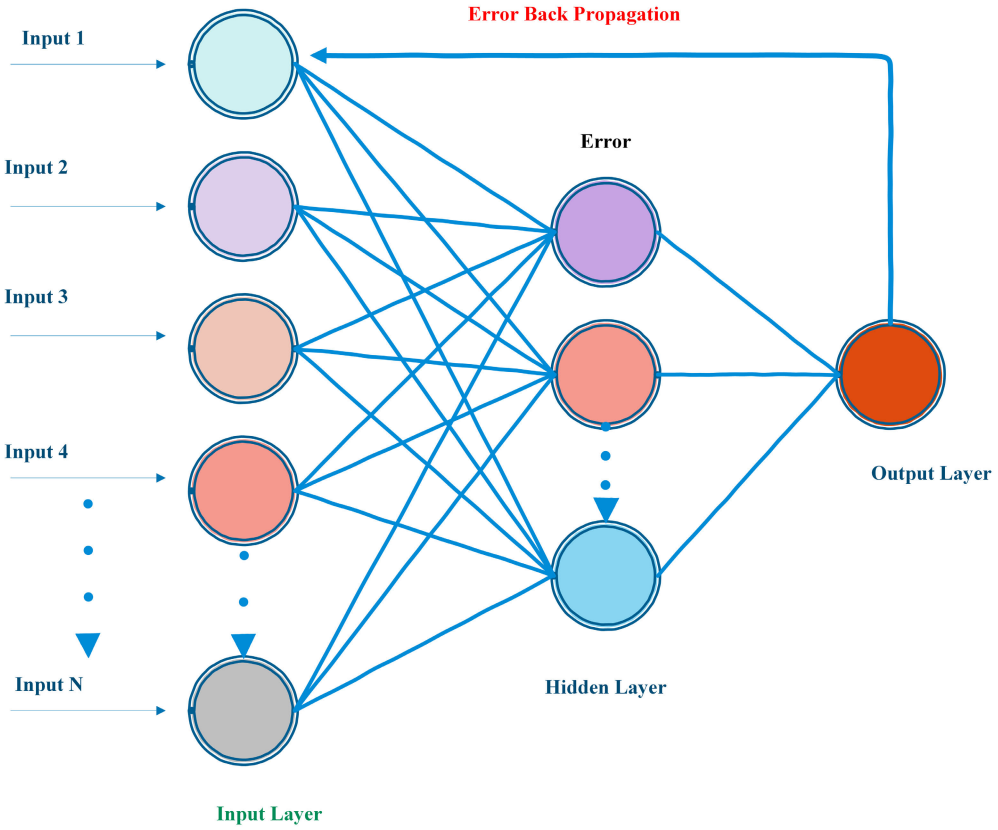

FIGURE 2. Architecture of ANN.

Hahn moments need input as a two-dimensional square matrix. So, Hahn polynomial can be expressed as given in Equation 13.

$$h_n^{u,v}(r,N) = (N+V-1)_n (N-1)_n \times \sum_{k=0}^{n} (-1)^k \frac{(-n)_k (-r)_k (2N+u+v-n-1)_k}{(N+v-1)_k (N-1)_k} \frac{1}{k!} I \quad (13)$$

For the calculation of the Pochhammer symbol a protocol mentioned in [50] is used. Then an orthogonal normalized Hahn moments are computed. Therefore, normalized Hahn moments for the 2-D matrix are computed as given by Equation 14.

$$H_{ij} = \sum_{q=0}^{N-1} \sum_{p=0}^{N-1} \beta_{ij} h_i^{\widetilde{u,v}}(q,N)\, h_j^{\widetilde{u,v}}(p,N) \quad (14)$$

D. FEATURE VECTOR

The feature vector is n-dimension vector that contains numerical values as features of an object. As all attributes and features in the datasets are mostly not used, features that perform better roles in prediction are extracted from the dataset. The patterns of input data are used for learning the model and the distinct properties of those patterns are extracted as features. In the proposed model, 522 features are extracted. They are used for classifying or differentiating the input patterns. It gave a vector of features to help in prediction. The feature vector of n-dimension contains numerical values as features of an object. PRIM, RPRIM, APPIV, and RAAPIV have been used to extract novel features and apply various statistical moments such as raw moments, Hanh moments, and central moments. These moments are applied on huge n-dimensions data accordingly and provided significant properties. These feature vectors are further used to train various classifiers. Each classifier is rigorously tested based on well-known validation techniques such as self-consistency, cross-validation, jackknife testing, and independent testing.

IV. PREDICTION ALGORITHMS

In this article, we use three machine learning models for prediction i.e., ANN, RF, and SVM to predict cancer driver genes and non-cancer driver genes as discussed in the following subsections.

A. ARTIFICIAL NEURAL NETWORK (ANN)

ANN consists of interconnected layers of neurons and can be used for the prediction of cancer driver genes [51]. The architecture of the back-propagation network is shown in Figure 2. ANN method is used based on the network that transmits the feed and uses an inverse spreading algorithm

to reduce the number of errors. The layer is related to vector design and has a hidden layer that gets the number of neurons from the input layer and then creates a processing section for the entire network. The ANN launch section collects larger and larger records in addition to non-standard values comprised of a three-stage continuous flow and an inverse error [52] as given by Equation 15.

$$O_m = f\left(\sum_{y=1}^{h} W_{ym} \times f\left(\sum_{x=1}^{k} W_{xy} I_a\right)\right) \tag{15}$$

The input and hidden layers consist of k and h neurons respectively. Each neuron computes the output denoted by O_m. For any node with input I_a, the weight of the edge connecting random node x to node y is denoted with W_{xy}. Whereas W_{ym} represents the weight of the node y connected to the neuron of the random output layer m, the function f in the equation. It is determined to be the classic sigma function that shows neuron activation as described in Equation 16.

$$f(x) = \frac{1}{\left(1 + e^{-x}\right)} \tag{16}$$

In every training pattern, the output units and the target output are compared. In every training pattern, the generated and targeted units are compared. E denotes the error rate which can be calculated using Equation 17.

$$E = 0.5 \sum_{i=1}^{o} (O_i - P_i) \tag{17}$$

where O_i is the target version and P_i is the latest count of the network. The gradient slope is used to reduce errors. The error design of the release layer is returned to the output layer. The layer of each weight is represented by a vector V. The recovery process chooses a different vector such as V to reduce errors. This continues periodically until the assembly is complete, as described in Equation 18.

$$V(t+1) = V(t) + \Delta(Vt) \tag{18}$$

Change in weight at time t+1 is computed using Equation 19. Where η is a positive constant and represents the rate of learning with the value between 0 and 1.

$$\Delta V = \eta\left(-\frac{\partial E}{\partial W}\right) | V = V(t) \tag{19}$$

Equation 20 is used to express the change in weights.

$$\Delta V_{u,v} = \eta\left(-\frac{\partial E}{\partial W_{u,v}}\right) \tag{20}$$

Here, $\Delta V_{u,v}$ indicates minimum E weight between u^{th} and v^{th} neurons in the i^{th} iteration. This procedure is also applied when crossing the front and back entry marks. This procedure is also applied when crossing the front and back entry marks. It is a lightweight system with low memory consumption used for training ANN. The target in the networks is usually to minimize the Mean Square Error (MSE) as given by Equation 21.

$$MSE = \frac{1}{m_a\, o} \sum_{m=1}^{m} \sum_{n=1}^{o} (P_{mn} - O_{mn}) \tag{21}$$

where P and O show the actual output and the output neurons respectively. In Equation 21, O_{mn} P_{mn} represent the predicted and the observed values respectively. In this article, the database included 763 cancers positive and 1805 negative genes. The Feature Input Matrix (FIM) is designed for the driver genes. Each FIM string represents a data model. Again, the Expected Output Matrix (EOM) is formed to confirm the corresponding FIM element class as positive or negative. ANN is trained using the input matrix FIM and the expected output matrix EOM. FIM is provided as an entry-level training module where EOM is used to calculate the error by backward propagating.

B. RANDOM FOREST (RF)

RF is used for regression and classification problems [53]. Therefore, the RF classifier is used here to predict both cancer-driver and non-cancer-driver genes. In the first step, the complete data is converted into a decision tree [54]. The class is predicted for each tree using the classifier. The generated feature input matrix of two data samples is used by the ANN algorithm. The model is trained on this data for prediction and the accuracy is calculated. The predicted class with the highest votes predicts the models as shown in Figure 3.

C. SUPPORT VECTOR MACHINE (SVM)

SVM is a machine learning classifier used for classification [55] and regression-related problems [56]. The primary objective of SVM is to identify a hyperplane in N-dimensional space, where N is the number of features that can be used to categorize a point. The hyperplane is a decision boundary used for the data points classification. The data points on opposite ends of the hyperplane are classified into separate classes [57]. The points on opposite sides of the hyperplane represent different classes i.e., class A and B as shown in Figure 4.

V. RESULTS AND DISCUSSION

This section presents experimental evaluation of the proposed framework. The experiments were run on an Intel (R) Core i7-7500 with two CPUs@2.70 GHz, 8 GB of memory, with 64-bit Ubuntu 18.04 operating system. Experimental datasets are obtained from NCBI web portal [46]. The dataset is freely available with verified statistics. Python 3.7 with Numpy and Sklearn are used for the implementation of the proposed model. Initially, pre-processing was done on sequencing data, the novel features were extracted and saved in the CSV file. Then, the CSV file is used as input to the classification models

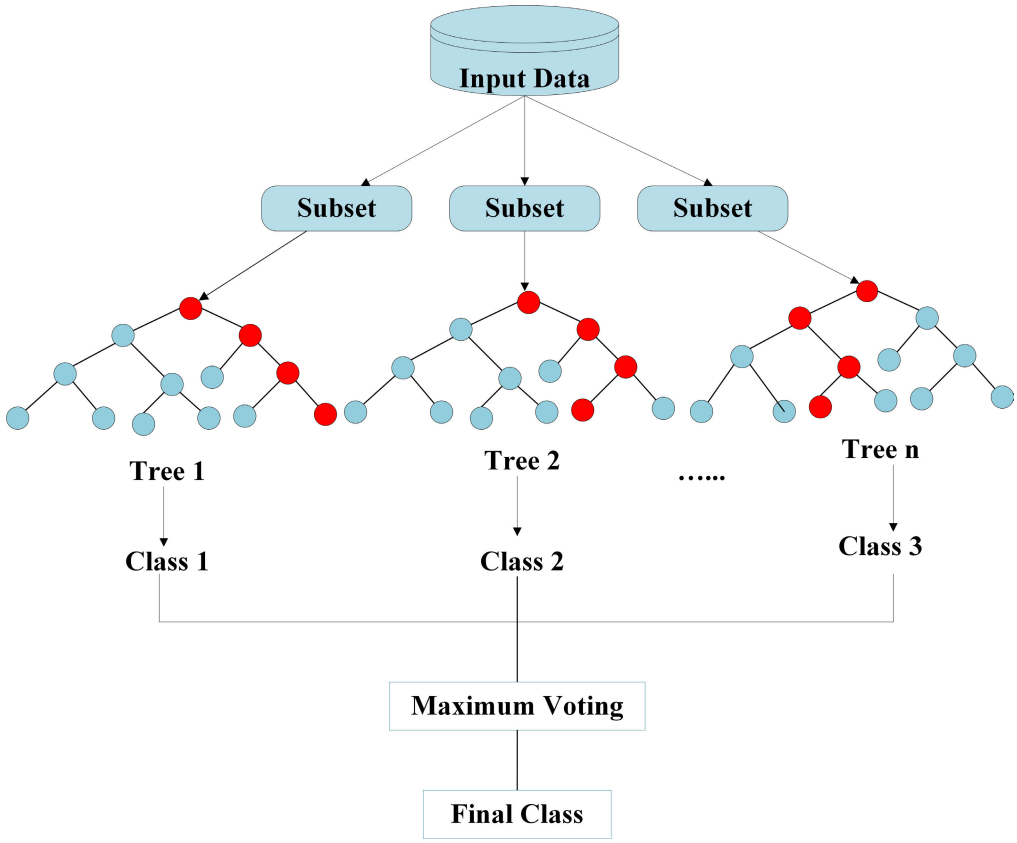

FIGURE 3. Architecture of RF.

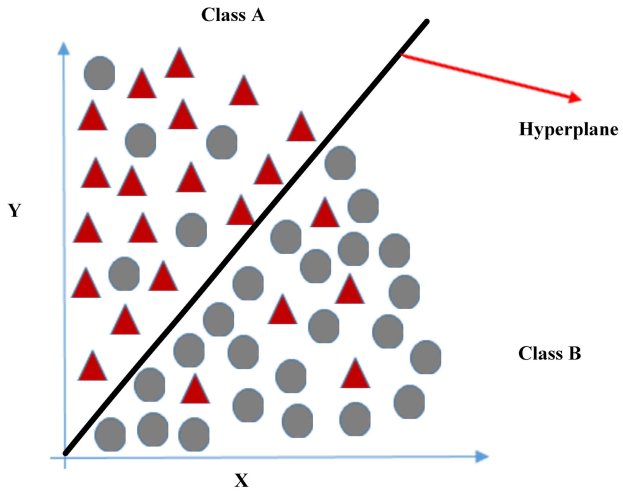

FIGURE 4. Architecture of support vector machine.

for prediction. Finally, the results generated are stored in the output file and plotted.

A. MODEL EVALUATION

Classification tests are measured in terms of specificity, accuracy, and sensitivity. The performance of a model can be measured statistically using these parameters. Sensitivity and specificity measure different categories or classes in a given dataset. In case the model detects a driver gene it will be either true which is referred to as True Positive (TP) or false which is referred to as False Positive (FP). In case the model does not detect driver genes it may be true referred to as True Negative (TN) or false referred to as False Negative (FN). Accuracy is the measurement of how accurately the model predicts the categories. Mathew Correlation Coefficient (MCC) is used for measuring the model stability. The measurement parameters as shown in Equations 22, 23, 24, and 25 are used for the evaluation of the proposed model and comparative analysis.

$$Specificity(SP) = \frac{TN}{(TN + FP)} \tag{22}$$

$$Sencitivity(SN) = \frac{TN}{(TP + FN)} \tag{23}$$

$$Accuracy(Acc) = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100 \tag{24}$$

$$MCC = \frac{(TP)\,(TN) - (FP)(FN)}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}} \tag{25}$$

VI. VALIDATION METHOD

Model testing is an important factor for the validation of the predicting model [58]. The proposed model is validated with

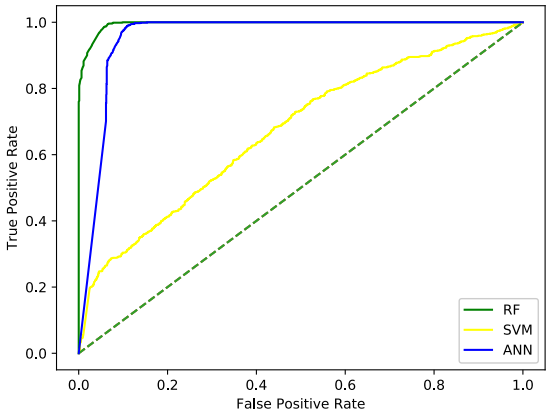

FIGURE 5. ROC graph of self-consistency.

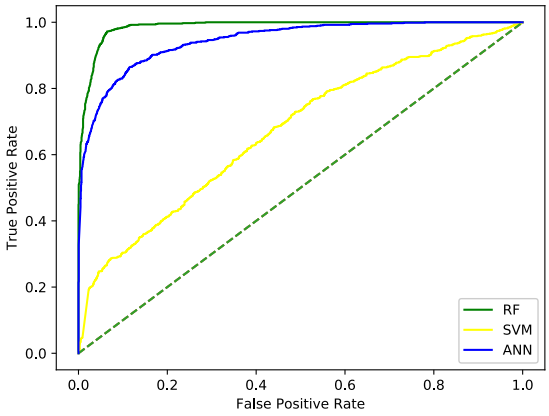

FIGURE 6. ROC graph of Jackknife.

the four tests i.e., jackknife testing, cross-validation, independent testing, and self-consistency. Following subsections present the detailed results.

A. SELF-CONSISTENCY

One of the simplest and most obvious tests is the self-consistency test. A trained model is simply evaluated using the training set of data. It serves as a simple yet effective benchmark for assessing the learning ability of a model. Self-consistency test was performed on both positive and negative genes using the datasets used for training the model. Table 1 presents the results of the self-consistency test and Figure 5 shows the ROC graph of the same test. The results analysis show that the RF classifier performs better than ANN and SVM classifier.

B. JACKKNIFE TESTING

Jackknife testing is the strictest testing method. Each iteration sees the removal of a sample while the algorithm is being trained on the remaining samples. The model is assessed using the missing sample after sufficient training. For each data sample, this process is repeated. Since N is the size of the testing set, this is carried out N times. Each cycle's testing data sample is unique; therefore, each sample test is conducted precisely once. Although this process is the strictest, it also takes the longest time [59]. The confusion matrix that the model develops after effectively training and testing include TP, FP, TN, and FN values that are used to compute accuracy for a particular instance ρ_j. The mean accuracy for all the instances is computed as depicted in Equation 26.

$$\bar{\rho}_j = \frac{1}{n}\sum_{j=1}^{m} \rho_j \tag{26}$$

where $\bar{\rho}_j$ represents the cumulative accuracy of the proposed model. The cumulative accuracy obtained for this test remains unique as the sample is tested once [60], [61]. Due to uniqueness, the results remain invariant and the test is considered more credible. Table 1 shows results obtained from jackknife using different classifiers. The results show that the accuracy of RF, ANN, and SVM is 91.3%, 88.4%, and 69.2% respectively. The ROC curve graph for jackknife testing validation is shown in Figure 6.

C. CROSS-VALIDATION

When testing requires anonymized data, but none is readily available, the cross-validation technique is utilized. The dataset is divided into multiple segments randomly which leads to rigorous testing [62]. In this technique, each partition is converted to disjoint from all other partitions. The training is performed on the data without considering the selected partition. After training is completed, the selected partition is used to test the model. The accuracy of the model is calculated at each iteration and the mean of all the obtained results is used for the cross-validation test. In the evaluation of the proposed model, 5-fold and 10-fold cross-valuations were used using the benchmark dataset. Table 1 shows the results of the cross-validation test. Figures 7 and 8 illustrate the ROC curve graph for 5-folds and 10 folds respectively.

D. INDEPENDENT TESTING

After training the model, independent testing is carried out using test data. To avoid ambiguity in the results the training and testing data are kept with similar ratios. Table 1 shows the results of the proposed model for independent testing in terms of accuracy on different classifiers i.e., RF, ANN, and SVM. The accuracy achieved by RF, ANN, and SVM is 95.4%, 93.0%, and 66.6% respectively. The ROC curve graph for independent testing is shown in Figure 9.

VII. COMPARATIVE ANALYSIS OF CLASSIFIERS

Comparative results of RF, ANN, and SVM are shown in Table 2. The analysis shows that in both cases i.e., cancer and non-cancer driver genes, RF achieves the highest accuracy approximately 95.8%, compared to ANN and SVM classifiers. ANN classifier shows 93.0% accuracy which is better than SVM while SVM shows a 69.2% accuracy. Figure 10 shows the graphical representation of comparative results.

TABLE 1. Comparative results analysis of RF, ANN and SVM with different validation methods.

Evaluation	Predictor	TN	FP	FN	TP	Acc (%)	SP (%)	SN (%)	MCC
Self-Consistency	RF	704	59	46	1697	95.8	92.3	97.4	0.90
	ANN	698	65	135	1608	92	91.5	92.3	0.81
	SVM	295	468	303	1440	69.2	38.7	82.6	0.22
5-Fold Cross validation	RF	346	417	19	1724	82.6	45.3	98.9	0.57
	ANN	531	232	57	1686	88.4	69.6	96.7	0.72
	SVM	295	468	303	1440	69.2	38.7	82.6	0.22
10-Fold Cross Validation	RF	620	143	32	1711	93.0	81.2	98.2	0.84
	ANN	531	232	57	1686	88.4	69.6	96.7	0.72
	SVM	295	468	303	1440	69.3	38.7	82.6	0.22
Jack-knife	RF	568	195	23	1720	91.3	74.4	98.7	0.79
	ANN	531	232	57	1686	88.4	69.6	96.7	0.72
	SVM	295	468	303	1440	69.2	38.7	82.6	0.22
Independent	RF	210	30	2	510	95.4	87.4	99.6	0.90
	ANN	200	37	15	500	93.0	84.4	97.1	0.83
	SVM	81	156	110	405	64.6	64.6	34.2	0.22

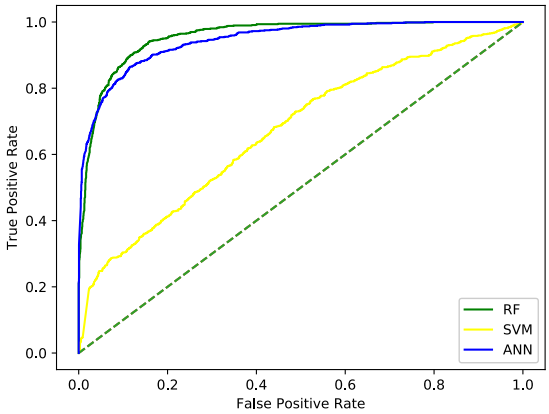

FIGURE 7. ROC graph of 5-fold cross validation.

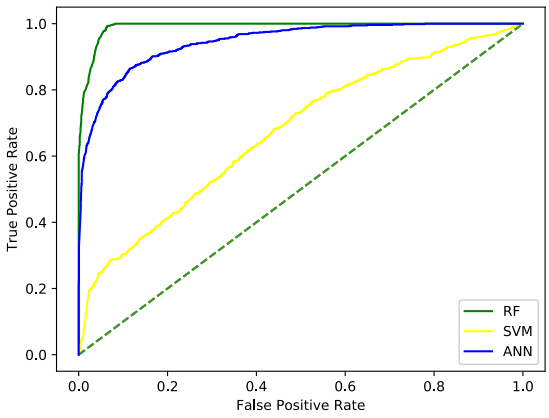

FIGURE 8. ROC graph of 10-fold cross validation.

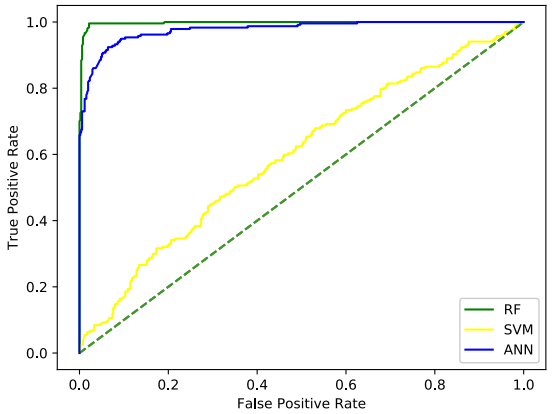

FIGURE 9. ROC graph of independent testing.

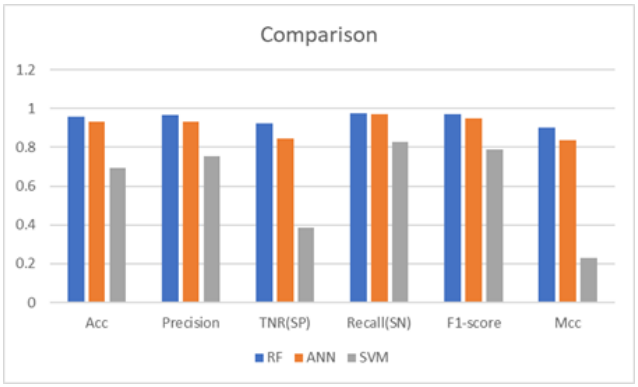

FIGURE 10. Overall performance of RF, ANN, and SVM.

The results show that RF performs better than ANN and SVM. RF shows 95.8% accuracy for cancer driver genes, ANN shows 93% accuracy while the accuracy of SVM is 69.2% for cross-validation and self-consistency tests. The performance of the model is measured using classification scores. Artificial dataset with heavy class imbalance does not yield better results and this type of validation becomes less effective. In such cases, Area Under the Curve (AUC) integrated with ROC is also a critical parameter in evaluating the performance of the classification models [63].

ROC curve is a plot of True Positive Rate (TPR) against False Positive Rate (FPR). It tells how much the model can distinguish between the classes. Higher accuracy means the model distinguishes the classes more accurately. The performance of the model is considered by looking at the AUC in a plot with TPR and FPR [64]. If the model distinguishes the classes accurately the accuracy tends toward 1 and tends

TABLE 2. Overall performance of RF, ANN and SVM.

Algorithm	Accuracy	Precision	TNR(SP)	Recall (SN)	F1-score	MCC
RF	0.95	0.96	0.92	0.97	0.96	0.90
ANN	0.93	0.93	0.84	0.97	0.95	0.83
SVM	0.69	0.75	0.38	0.82	0.78	0.22

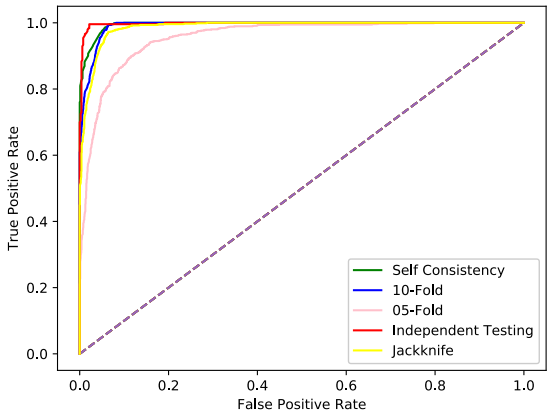

FIGURE 11. ROC graph of comparative analysis of RF.

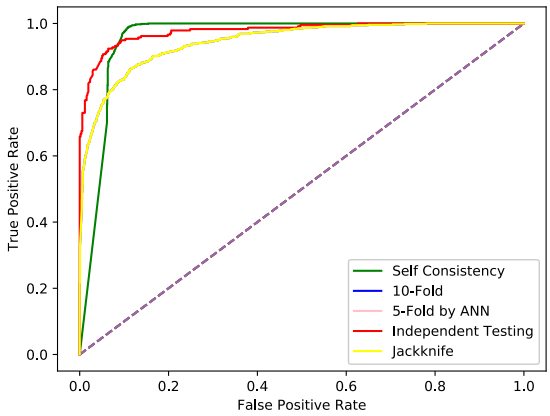

FIGURE 12. ROC graph of comparative analysis of ANN.

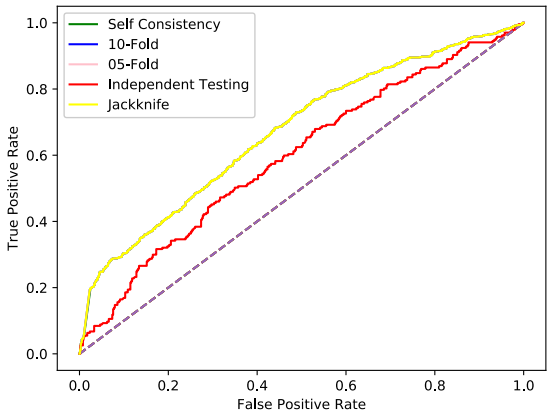

FIGURE 13. ROC graph of comparative analysis of SVM.

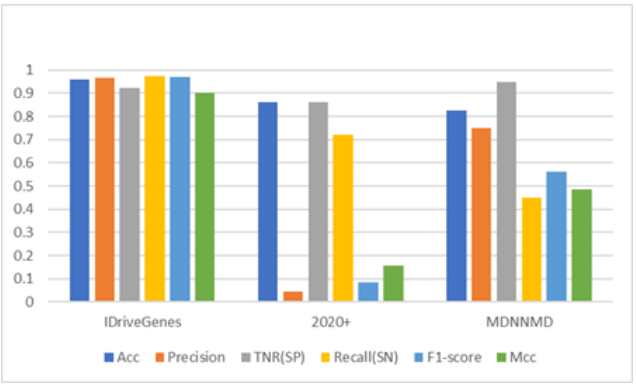

FIGURE 14. Comparisons of existing methods.

to zero in case the classes are distinguished less accurately. The effectiveness of a classifier is measured using various measurement techniques. To validate the performance of the proposed model the different classifiers are used, and the results are presented using ROC curve. Figures 11, 12, and 13 shows the results of RF, ANN, and SVM classifiers respectivley. Results show that the area of RF is near to1, which indicates the RF has better measure of separating the classes.

VIII. COMPARISON WITH EXISTING METHODS

The proposed model IDriveGenes is compared with the state-of-the-art models i.e., 20/20+ and MDNNMD. The 20/20+ model was proposed to differentiate driver genes from passenger mutation in cancer [30]. It utilized the RF tree trained on known cancer driver genes to recognize cohort-level cut-offs that fit the said type of identification. Such a method requires previous molecular knowledge for the alteration. MDNNMD [26] proposed for breast cancer diagnosis from multi-dimensional data includes gene expression profile data and CNA profile data. The method selects features by applying the mRMR [17] feature selection method, which effectively reduces data dimensionality without losing important information. The said method selects 400 genes and 200 genes from gene expression profile data and CNA profile data. To show the significance of the proposed model, various predictive performance measures such as accuracy, precision, SP, SN (Recall), F1-measure score, and MCC are measured. The IDriveGenes yields 95.8% accuracy which gives correct identifications from the total dataset. Both existing methods achieve 86% and 82% accuracy respectively which is less than the proposed model. The proposed model obtained 92.2% SP and 97.3% SN which is better than existing models. The precision of the IDriveGenes is 96.6% which computes several correct positive predictions from a total number of positive predictions. F1-score of IDriveGenes is almost 97%. MCC of the IDriveGenes is 90% which is higher than the existing techniques. Table 3

TABLE 3. Comparative results analysis of proposed model with existing.

Model	Algorithm	Accuracy	Precision	TNR(SP)	Recall (SN)	F1-score	MCC
IDriveGenes	RF	0.95	0.96	0.92	0.97	0.96	0.90
20/20+	RRF	0.86	0.044	0.86	0.72	0.084	0.15
MDNNMD	DNN	0.82	0.74	0.95	0.45	0.56	0.48

shows the detailed comparison and Figure 14 illustrates the comparison in graphical form of the IDriveGenes with existing methods. It can be concluded that the results shown by the IDriveGenes are better than 20/20+ and MDNNMD. Moreover, the proposed model takes less time and consumes less memory for training as compared to existing approaches; however, it needs more time to extract robust features from the NGS dataset. The execution time of the training model of RF is less than ANN. RF is also less susceptible to overfitting. Accuracy of a model is strongly influenced by the robustness of the feature extraction technique. More relevant features to the composition and sequence of the primary structures are extracted during feature extraction. Correct feature extraction leads to better results for a model. The proposed model yields result in less time and cost from a given sequence. The effectiveness of the model in identifying driver genes is evident from the results.

IX. CONCLUSION

Cancer is caused by cell proliferation. Cancer genes proliferate following mutation. The definition of such cells can help with treatment and, in some cases, even the cure of the ailment. This study suggests a secure in-silico method for using classifiers to find cancer genes. To obtain qualities out from reference dataset, an optimal feature drilling approach was applied. The following feature vectors were used to build classifiers which including ANN, SVM, and RF. Once models have been fully trained, test techniques including Jackknife testing, self-consistency, k-fold cross validation, and independent testing are rigorously tested. For cancer driver genes, the Random Forest classifier had 95% accuracy, the ANN had 92% accuracy, and the SVM had 69% accuracy. On the contrary hand, the suggested structure outperformed the current approaches in every way.

ACKNOWLEDGMENT

This work is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R125), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] M. Sardaraz and M. Tahir, "SCA-NGS: Secure compression algorithm for next generation sequencing data using genetic operators and block sorting," *Sci. Prog.*, vol. 104, no. 2, Apr. 2021, Art. no. 003685042110232.

[2] A. Alourani, M. Tahir, M. Sardaraz, and M. S. Khan, "Knowledge-based framework for selection of genomic data compression algorithms," *Appl. Sci.*, vol. 12, no. 22, p. 11360, Nov. 2022.

[3] J. Reimand and G. D. Bader, "Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers," *Mol. Syst. Biol.*, vol. 10, no. 8, p. 637, Aug. 2014.

[4] L.-H. Wang, C.-F. Wu, N. Rajasekaran, and Y. K. Shin, "Loss of tumor suppressor gene function in human cancer: An overview," *Cellular Physiol. Biochem.*, vol. 51, no. 6, pp. 2647–2693, 2018.

[5] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Res.*, vol. 40, no. 19, pp. 9379–9391, Oct. 2012.

[6] R. Beroukhim, "Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 50, pp. 20007–20012, 2007.

[7] S. Samaddar, R. Sinha, and R. K. De, "A model for distributed processing and analyses of NGS data under map-reduce paradigm," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 827–840, May 2019.

[8] K. Brennan, G. Offiah, E. A. McSherry, and A. M. Hopkins, "Tight junctions: A barrier to the initiation and progression of breast cancer?" *J. Biomed. Biotechnol.*, vol. 2010, pp. 1–16, Nov. 2010.

[9] N. E. Davidson, S. A. Armstrong, L. M. Coussens, M. R. Cruz-Correa, R. J. DeBerardinis, J. H. Doroshow, M. Foti, P. Hwu, T. W. Kensler, and M. Morrow, "AACR cancer progress report 2016," *Clin. Cancer Res., Off. J. Amer. Assoc. Cancer Res.*, vol. 22, pp. S1–S137, Oct. 2016.

[10] P. K. Kreeger and D. A. Lauffenburger, "Cancer systems biology: A network modeling perspective," *Carcinogenesis*, vol. 31, no. 1, pp. 2–8, Jan. 2010.

[11] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, Jan. 2006.

[12] C. Liu, Y. Dai, K. Yu, and Z.-K. Zhang, "Enhancing cancer driver gene prediction by protein-protein interaction network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 4, pp. 2231–2240, Jul. 2022.

[13] V. V. H. Pham, L. Liu, C. Bracken, G. Goodall, J. Li, and T. D. Le, "Computational methods for cancer driver discovery: A survey," *Theranostics*, vol. 11, no. 11, pp. 5553–5568, 2021.

[14] X. Shi, H. Teng, L. Shi, W. Bi, W. Wei, F. Mao, and Z. Sun, "Comprehensive evaluation of computational methods for predicting cancer driver genes," *Briefings Bioinf.*, vol. 23, no. 2, Mar. 2022, Art. no. bbab548.

[15] R. Andrades and M. Recamonde-Mendoza, "Machine learning methods for prediction of cancer driver genes: A survey paper," *Briefings Bioinf.*, vol. 23, no. 3, May 2022, Art. no. bbac062.

[16] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 971–989, Sep. 2016.

[17] M. Gutkin, R. Shamir, and G. Dror, "SlimPLS: A method for feature selection in gene expression-based disease classification," *PLoS ONE*, vol. 4, no. 7, Jul. 2009, Art. no. e6416.

[18] Cancer Genome Atlas Research Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, p. 1061, 2008.

[19] J. Barretina, G. Caponigro, N. Stransky, and K. Venkatesan, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012.

[20] L. Ding, G. Gets, and R. K. Wilson, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, no. 7216, pp. 1069–1075, Oct. 2008.

[21] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," *Genome Res.*, vol. 22, no. 2, pp. 398–406, Feb. 2012.

[22] S. Jones, "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses," *Science*, vol. 321, no. 5897, pp. 1801–1806, Sep. 2008.

[23] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[24] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000.

[25] P. T. Tran, A. C. Fan, P. K. Bendapudi, S. Koh, K. Komatsubara, J. Chen, G. Horng, D. I. Bellovin, S. Giuriato, C. S. Wang, J. A. Whitsett, and D. W. Felsher, "Combined inactivation of MYC and K-Ras oncogenes reverses tumorigenesis in lung adenocarcinomas and lymphomas," *PLoS ONE*, vol. 3, no. 5, May 2008, Art. no. e2125.

[26] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 841–850, May 2019.
[27] J. Tan, J. H. Hammond, D. A. Hogan, and C. S. Greene, "ADAGE analysis of publicly available gene expression data collections illuminates *Pseudomonas aeruginosa*-host interactions," *BioRxiv*, Nov. 2015, Art. no. 30650.
[28] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[29] P. Luo, Y. Ding, X. Lei, and F.-X. Wu, "DeepDriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks," *Frontiers Genet.*, vol. 10, p. 13, Jan. 2019.
[30] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, "Evaluating the evaluation of cancer driver genes," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 50, pp. 14330–14335, Dec. 2016.
[31] M. E. Sana, M. Iascone, D. Marchetti, J. Palatini, M. Galasso, and S. Volinia, "GAMES identifies and annotates mutations in next-generation sequencing projects," *Bioinformatics*, vol. 27, no. 1, pp. 9–13, Jan. 2011.
[32] Y. Han, J. Yang, X. Qian, W.-C. Cheng, S.-H. Liu, X. Hua, L. Zhou, Y. Yang, Q. Wu, P. Liu, and Y. Lu, "DriverML: A machine learning algorithm for identifying driver genes in cancer sequencing studies," *Nucleic Acids Res.*, vol. 47, no. 8, p. e45, May 2019.
[33] M. Rahimi, B. Teimourpour, and S.-A. Marashi, "Cancer driver gene discovery in transcriptional regulatory networks using influence maximization approach," *Comput. Biol. Med.*, vol. 114, Nov. 2019, Art. no. 103362.
[34] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: Application to cancer genomics," *Nucleic Acids Res.*, vol. 39, no. 17, p. e118, Sep. 2011.
[35] L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, and N. López-Bigas, "OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations," *Genome Biol.*, vol. 17, no. 1, p. 128, Dec. 2016.
[36] A. Gonzalez-Perez, J. Deu-Pons, and N. Lopez-Bigas, "Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation," *Genome Med.*, vol. 4, no. 11, p. 89, 2012.
[37] O. Collier, V. Stoven, and J.-P. Vert, "LOTUS: A single- and multitask machine learning algorithm for the prediction of cancer driver genes," *PLOS Comput. Biol.*, vol. 15, no. 9, Sep. 2019, Art. no. e1007381.
[38] M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz, "Discovery and saturation analysis of cancer genes across 21 tumour types," *Nature*, vol. 505, no. 7484, pp. 495–501, Jan. 2014.
[39] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, "MuSiC: Identifying mutational significance in cancer genomes," *Genome Res.*, vol. 22, no. 8, pp. 1589–1598, Aug. 2012.
[40] T. Davoli, A. W. Xu, K. E. Mengwasser, L. M. Sack, J. C. Yoon, P. J. Park, and S. J. Elledge, "Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome," *Cell*, vol. 155, no. 4, pp. 948–962, Nov. 2013.
[41] Y. Fu et al., "A computational framework for prioritizing noncoding regulatory variants in cancer," *Cancer Res.*, vol. 75, no. 15_Suppl., p. 4854, 2015.
[42] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, Sep. 2010.
[43] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas, "IntOGen-mutations identifies cancer drivers across tumor types," *Nature Methods*, vol. 10, no. 11, pp. 1081–1082, Nov. 2013.
[44] H. Carter, J. Samayoa, R. H. Hruban, and R. Karchin, "Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM)," *Cancer Biol. Therapy*, vol. 10, no. 6, pp. 582–587, Sep. 2010.
[45] M. D. M. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael, "Simultaneous identification of multiple driver pathways in cancer," *PLoS Comput. Biol.*, vol. 9, no. 5, May 2013, Art. no. e1003054.
[46] *National Center for Biotechnology Information*. Accessed: May 10, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/genome
[47] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: A web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010.
[48] Y. D. Khan, E. Alzahrani, W. Alghamdi, and M. Z. Ullah, "Sequence-based identification of allergen proteins developed by integration of PseAAC and statistical moments via 5-step rule," *Current Bioinf.*, vol. 15, no. 9, pp. 1046–1055, Nov. 2020.
[49] A. H. Butt, N. Rasool, and Y. D. Khan, "Predicting membrane proteins and their types by extracting various sequence features into Chou's general PseAAC," *Mol. Biol. Rep.*, vol. 45, no. 6, pp. 2295–2306, Dec. 2018.
[50] M. A. Akmal, N. Rasool, and Y. D. Khan, "Prediction of N-linked glycosylation sites using position relative features and statistical moments," *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0181966.
[51] S. Muthusamy, L. P. Manickam, V. Murugesan, C. Muthukumaran, and A. Pugazhendhi, "Pectin extraction from helianthus annuus (sunflower) heads using RSM and ANN modelling by a genetic algorithm approach," *Int. J. Biol. Macromolecules*, vol. 124, pp. 750–758, Mar. 2019.
[52] L. Jiang, J. Zhang, P. Xuan, and Q. Zou, "BP neural network could help improve pre-miRNA identification in various species," *BioMed Res. Int.*, vol. 2016, pp. 1–11, Aug. 2016.
[53] C. Kathuria, D. Mehrotra, and N. K. Misra, "Predicting the protein structure using random forest approach," *Proc. Comput. Sci.*, vol. 132, pp. 1654–1662, Jan. 2018.
[54] H. Cao, S. Bernard, R. Sabourin, and L. Heutte, "Random forest dissimilarity based multi-view learning for Radiomics application," *Pattern Recognit.*, vol. 88, pp. 185–197, Apr. 2018.
[55] H. M. Lalabadi, M. Sadeghi, and S. A. Mireei, "Fish freshness categorization from eyes and gills color features using multi-class artificial neural network and support vector machines," *Aquacultural Eng.*, vol. 90, Aug. 2020, Art. no. 102076.
[56] D. Willsch, M. Willsch, H. D. Raedt, and K. Michielsen, "Support vector machines on the D-Wave quantum annealer," *Comput. Phys. Commun.*, vol. 248, Mar. 2020, Art. no. 107006.
[57] Y. Yang, J. Li, and Y. Yang, "The research of the fast SVM classifier method," in *Proc. 12th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2015, pp. 121–124.
[58] F. Javed and M. Hayat, "Predicting subcellular localization of multi-label proteins by incorporating the sequence features into Chou's PseAAC," *Genomics*, vol. 111, no. 6, pp. 1325–1332, Dec. 2019.
[59] W. Hussain, Y. D. Khan, N. Rasool, S. A. Khan, and K.-C. Chou, "SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins," *J. Theor. Biol.*, vol. 468, pp. 1–11, May 2019.
[60] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "DPP-PseAAC: A DNA-binding protein prediction model using Chou's general PseAAC," *J. Theor. Biol.*, vol. 452, pp. 22–34, Sep. 2018.
[61] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.
[62] X. Cui, Z. Yu, B. Yu, M. Wang, B. Tian, and Q. Ma, "UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components," *Chemometrics Intell. Lab. Syst.*, vol. 184, pp. 28–43, Jan. 2019.
[63] A. C. J. W. Janssens and F. K. Martens, "Reflection on modern methods: Revisiting the area under the ROC curve," *Int. J. Epidemiol.*, vol. 49, no. 4, pp. 1397–1403, Aug. 2020.
[64] J. V. Carter, J. Pan, S. N. Rai, and S. Galandiuk, "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves," *Surgery*, vol. 159, no. 6, pp. 1638–1645, Jun. 2016.

YASIR ALI received the master's degree in computer science from the Department of Computer Science, COMSATS University Islamabad, Attock Campus. He is currently a Lecturer with the Department of Computer Science, Sir Syed Case Institute of Technology, Islamabad. His research interests include machine learning, optimization techniques, and healthcare applications.

MUHAMMAD SARDARAZ received the master's degree in computer science from Foundation University Islamabad and the Ph.D. degree in computer science from Iqra University Islamabad, Pakistan, in 2016. He was a Lecturer with the Department of Computer Science, University of Wah, Wah Cantonment. He is currently an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Attock Campus, Pakistan. His research interests include cloud computing, cluster and grid computing, machine learning, and bioinformatics.

MUHAMMAD TAHIR received the Ph.D. degree in computer science from the Department of Computing and Technology, Iqra University, Islamabad, Pakistan, in 2016. He was a Lecturer with the Department of Computer Science, University of Wah, Wah Cantonment. He is currently an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Attock Campus, Pakistan. His research interests include parallel and distributed computing, Hadoop MapReduce framework, the Internet of Things, and machine learning.

HELA ELMANNAI received the Ph.D. degree in information technology from Sup'Com, Tunisia. She is currently an Assistant Professor with the Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Saudi Arabia. Her research interests include artificial intelligence, networking, blockchain, and engineering applications.

MONIA HAMDI received the B.E. degree in information technology from Telecom SudParis, Paris-Saclay University, Gif-sur-Yvette, France, in 2008, the M.Sc. degree in telecommunications and networks from the Institut National Polytechnique, Toulouse, France, in 2008, and the Ph.D. degree in computer science from the University Rennes 1, Rennes, France, in 2012. From 2012 to 2017, she was an Assistant Professor with the Higher Institute of Computer Science and Multimedia, Gabes University, Gabes, Tunisia. From March 2015 to August 2015, she was a Visiting Researcher with the Department of Science and Technology, Linköping University, Linköping, Sweden. She is currently an Assistant Professor with the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include cooperative communications networks and self-organizing wireless networks. She was a member of the technical program committees of several conferences and journals, such as IEEE International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, IEEE International Wireless Communications and Mobile Computing Conference, and *IET Communications*.

AMEL KSIBI received the B.S., M.S., and Ph.D. degrees in computer engineering from the National School of Engineering of Sfax (ENIS), University of Sfax, Tunisia, in 2008, 2010, and 2014, respectively. She spent three years at ENIS as a Teaching Assistant, before joining the Higher Institute of Computer Science and Multimedia Gabes (ISIMG) as a permanent Lecturer, in 2013. She joined the Computer Science Department, Umm Qura University (UQU), as an Assistant Professor, in 2014. She joined Princess Nourah Bint Abdulrahman University, in 2018, where she is currently an Assistant Professor with the Department of Information Systems, College of Computer Sciences and Information. Her research interests include computer vision, image processing, deep learning, information retrieval, lifelogging and well-being, smart education, smart agriculture, and sustainable environment.