This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3109069, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Analysis of Transcriptomic microRNA using Text Mining Methods: An Early Detection of Multiple Sclerosis Disease

Nehal M. Ali[1], Member, IEEE, Mohamed Shaheen[2], Mai S. Mabrouk[3], and Mohamed A. Aborizka[1]

1 College of Computing and Information Technology, Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt
2 College of Computing and Information Technology, Arab Academy for Science Technology and Maritime Transport, Alexandria, Egypt.
3 Department of Biomedical Engineering, Faculty of Computer Science, Misr University for Science and Technology, Cairo, Egypt.

Corresponding author: Nehal M. Ali (e-mail: nehal_ali@ieee.org)

ABSTRACT Multiple sclerosis is an autoimmune disease that causes psychological impacts and severe physical disabilities, including motor disabilities and partial blindness. This work introduces an early detection method for multiple sclerosis disease by analyzing transcriptomic microRNA data. By transforming this phenotype classification problem into a text mining problem, multiple sclerosis disease biomarkers can be obtained. To our knowledge, text mining methods have not been introduced previously in transcriptomic data analysis of multiple sclerosis disease. Hence, this work presents a complete predictive model by combining consecutive transcriptomic data preprocessing procedures, followed by the proposed *KmerFIDF* method as a feature extraction method and linear discriminant analysis for dimensionality reduction. Predictive machine learning methods can then be obtained accordingly. This study describes experimental work on a transcriptomic dataset of noncoding microRNA sequences denoted from relapsing-remitting multiple sclerosis patients before fingolimod treatment and after six consecutive months of treatment. The experimental results of the predictive methods with the proposed model report sensitivity, specificity, F1-score, and average accuracy scores of 96.4, 96.47, 95.6, and 97% with random forest, 92.89, 92.78, 93.2, and 94% with support vector machine and 91.95, 92.2, 93.1, and 94% with logistic regression, respectively. These promising results support the introduced model and the proposed KmerFIDF method in transcriptomic data analysis. Moreover, comparative experiments are conducted with two referenced studies. The obtained results show that the average reported accuracy scores of the proposed model outperform the referenced literature work.

INDEX TERMS KmerFIDF, miRNA analysis, Multiple Sclerosis, Machine Learning, text mining, Transcriptomic data

I. INTRODUCTION

Multiple sclerosis (MS) is an autoimmune disease that causes demyelination of the myelin sheath of nervous cells. Myelin is a layer that insulates the nervous cells that make up the central nervous system (CNS). This myelin sheath permits electrical impulses to be transmitted efficiently and rapidly along with nerve cells. If myelin is damaged, these impulses slow down, causing severe physical and psychological impacts [1] [2].
This demyelination of the myelin sheath impacts the signal conduction between nervous cells. Consequently, it causes disabilities, including partial or total blindness, double vision, muscle weakness, motor disabilities, and psychological problems [3].

In microbiology, RNA sequencing is a specific technology based on a sequencing technique that uses next-generation sequencing (NGS) to determine the presence and quantity of RNA within a given biological sample at a given moment. Analyzing RNA sequences promotes the ability to investigate mutations, SNPs, and changes in gene expression across populations or generations. Alternative gene spliced transcripts, gene fusion, or differences in gene expression in different groups or treatments can also be determined. Thanks to NGS, vast quantities of data on transcriptomes and human genomes are available to researchers [4].

In addition to mRNA transcripts, RNA sequencing can be obtained on several RNA populations to provide total RNA or small RNA, such as miRNA, tRNA, and ribosomal profiling. Moreover, RNA sequencing can also be used to obtain exon/intron boundaries and verify or alter previously annotated 5' and 3' gene boundaries [5].

miRNA is a short noncoding RNA molecule that binds to target mRNAs and regulates translational repression and gene silencing and is primarily expressed in all eukaryotic cells. Hence, miRNAs organize pathological and

physiological processes, such as development, proliferation, cell differentiation, apoptosis, and tumor growth. miRNA is considered as a type of noncoding sequence. Thus, it has not been regarded as informative for RNA expression analysis. Nevertheless, recent studies have commenced analyzing miRNAs in disease detection [6]. This work introduces a complete detailed model to analyze microRNA data through text mining methods for early MS biomarker detection. We studied an NGS dataset of relapsing-remitting MS patients before and after treatment with fingolimod for six consecutive months.

Thus, the primary contributions of this study are:

- Proposes a complete model of obtaining multiple sclerosis (MS) disease biomarkers by transforming the microRNA analysis problem into a text mining problem.
- To our knowledge, text mining methods have not been introduced previously in transcriptomic data analysis of MS disease.
- This work introduces kmerFIDF as a feature extraction method and examines it accordingly.
- The experimental results have an average accuracy score of 97.0 in disease biomarker detection.
- The proposed model results outperform literature results in obtaining MS biomarkers using the same dataset.
- The implications of this study indicate the high detection potential of MS biomarkers from transcriptomic data and the use of the introduced method compared to traditional EEG signal analysis.

II. BIOLOGICAL BACKGROUND

In molecular biology, gene expression is the procedure by which information from DNA is used to synthesize a functional gene product such as a protein. These procedures primarily involve two main phases: transcription and translation. In prokaryotes, transcription and translation are executed. On the other hand, a nuclear membrane separates the two processes in eukaryotes [7].

A. TRANSCRIPTION

Transcription is the process of replicating the information contained in a section of DNA to synthesize messenger RNA (mRNA). The enzymes that alleviate this replication process primarily involve RNA polymerase and transcription factors [8].

A transcription factor (TF) is a protein that binds to a specific DNA sequence to regulate the transcription rate of genetic information from DNA to messenger RNA. The primary role of TFs is to regulate genes to guarantee that they are expressed in the correct cell with the correct number and at the proper time.

In eukaryotes, the transcription process is carried out in the nucleus and involves three types of RNA polymerases. Each type requires specific transcription factors. Through these TFs and RNA polymerase, premRNA is produced as the primary transcript in eukaryotic cells. This pre-mRNA then goes through a series of alternations to synthesize mature mRNA [9].

Pre-mRNA must be prepared before the translation process commences. This preparation process includes adding a 5' cap and a poly-A tail to the pre-mRNA chain and then splicing.

5' Capping involves enzymatic reactions that add 7-methylguanosine (m7G) to the 5' end of pre-mRNA, which consequently protects the RNA from degradation by exonucleases. Then, the cap-binding complex heterodimer (CBC20/CBC80) binds with m7G. This binding supports mRNA export to the cytoplasm and protects the RNA from decapping[10].

A second alteration that pre-mRNA undergoes is 3' cleavage and polyadenylation. First, the pre-mRNA is cleaved, followed by a series of adenine (A) added to synthesize a poly(A) tail. This poly(A) tail preserves the RNA from degradation. Afterward, the poly(A) tail is bound by many poly (A)-binding proteins (PABPs), which are mandatory for mRNA export and translation reinitiation[11].

Additionally, RNA splicing is another major significant alternation in eukaryotic pre-mRNA. Most eukaryotic pre-mRNAs are made up of alternating segments of exons and introns.

The splicing process involves the spliceosome as an RNA-protein catalytically complex. Spliceosomes primarily catalyzes two transesterification reactions. The first removes an intron and releases it in the form of a lariat structure; the second splices the adjacent exons together. Moreover, under some circumstances, alternative splicing is applied. Some introns or exons are retained or removed within a mature messenger RNA (mRNA) [12].

Hence, a noncoding RNA (ncRNA) is a functional RNA molecule transcribed from DNA. The critical difference between it and the coding RNA is that it is not translated into proteins. Epigenetic-related ncRNAs primarily include Piwi-interacting RNA (piRNA), microRNA (miRNA), long noncoding RNAs (lncRNAs), and small interfering RNA (siRNA).

Their primary function is to regulate gene expression at both the transcriptional and post-transcriptional levels. These ncRNAs that appear to be involved in epigenetic processes can be divided into two main categories: short ncRNAs (<30 nucleotides) and long ncRNAs (>200 nucleotides). The three main classes of short noncoding RNAs are miRNAs, siRNAs, and piRNAs. Both the long and short ncRNA categories are involved in heterochromatin formation, histone modification, DNA methylation targeting, and gene silencing [13].

B. TRANSLATION

Following transcription of DNA into mRNA processes, translation procedures are performed to synthesize proteins. It takes place in ribosomes in the cytoplasm or endoplasmic reticulum. In the translation process, the messenger RNA

(mRNA) is first decoded into a specific amino acid chain (a polypeptide).
Afterward, the produced polypeptide folds into an active protein and performs its functions in the cell accordingly. The ribosome's role is to facilitate the decoding process by instigating the binding of complementary transfer RNA (tRNA) anticodon sequences to mRNA codons [8]. The translation process involves three main phases. The first phase is initiation, where the ribosome assembles around the target mRNA.

That is, the first tRNA is attached at the start codon. Afterward, the elongation phase pursues, where the tRNA transforms the amino acid into the tRNA corresponding to the next codon. Consequently, the ribosome then translocates to the next mRNA codon and eventually produces an amino acid chain. The final phase is termination, where the ribosome releases the polypeptide when a stop codon is reached [14].

III. LITERATURE REVIEW

Several studies have utilized machine learning methods to analyze medical data and introduced the analysis of several text mining methods. A study has investigated using active learning (AL) methods to label medical data. In addition, a comparative study of AL techniques using machine and deep learning methods was developed. The obtained results of the study showed that AL techniques have high potential in the cost reduction of manual labeling[15].
A second study analyzed the unstructured data of 124 journal articles employing text mining techniques. The results of the applied computational analysis and systematic manual of these articles illustrated the state and evolution of text mining applications and provided evidence-based recommendations regarding their future use[16].
Moreover, another report analyzed patients' biomedical data using the topic modeling technique for text mining through hybrid inverse document frequency and machine learning fuzzy k-means clustering method. The Calinsiki-Har-Abasz (CH) index internal validation method was used to evaluate the clustering performance. The reported results have shown the potential of enhancing the redundancy issue and determining topics from biomedical text documents[17].
Another study analyzed MS patients' data by processing the patients' progress notes and other clinical text in the electronic medical record (EMR) to identify MS phenotypes and classify them three classes: relapsing-remitting MS (RRMS), primary-progressive MS (PPMS), secondary-progressive MS (SPMS), or progressive-relapsing MS (PRMS) using natural into language processing (NLP) techniques given specific codes. The study introduced promising results of determining MS phenotypes from EMR. The article also reported a positive predictive value of 93.8% over studying a dataset of 145 EMR samples [18].
An NLP model was introduced to classify patients with systemic lupus erythematosus (SLE) disease from clinical notes. Bag-of-Words (BOWs) and Unified Medical Language System (UMLS) and Concept Unique Identifiers (CUIs) matrices were produced using NLP pipelines. Then, the following classifiers were applied to the produced matrices: shallow neural networks, random forest, naïve Bayes, and support vector machines with word2vec inversion. The results indicated that the random forest classifier reported the highest accuracy of 92.1% with CUI [19].
Another NLP method was developed to extract approximately 1000 MS-related terms that occurred significantly more frequently in MS patients' notes for signs and symptoms than in the controls'. Synonymous terms were manually clustered. Patients' notes were extracted from an outpatient clinic data warehouse, and signs and symptoms were mapped to Unified Medical Language System (UMLS) terms using MedLEE. A naïve Bayes classifier was obtained, and an accuracy of 90% was reported [20].

Phenotype prediction was conducted from a bacterial isolate by studying NGS of a prokaryotic dataset of bacterial isolates. The research team utilized high specifications hardware to develop software that uses text mining methods for feature extraction followed by regression analysis. The introduced method was validated against a dataset of 167 Klebsiella pneumonia isolates, 200 Pseudomonas aeruginosa isolates, and 459 Clostridium difficile isolates. The presented work has reported a prediction accuracy of 88% [21].

An automated workflow that analyzes shotgun sequencing metagenomics data was proposed. A user-friendly interface was introduced to facilitate interactions and integrated with other popular software such as FASTQC, MultiQC, Trimmomatic, and Diamond[35].
A stand-alone application for NGS data processing and retrieval was proposed, where the *Aspera high-speed* file transfer protocol was utilized to maintain transfer speed optimization. FASTQC software was used to evaluate the quality of raw sequence data. In addition, Trimmomatic was used for data trimming and alignment. The obtained results reported that the proposed tool has outperformed a similar processing speed tool [45].
Another classification system was conducted to classify phase 1 and two cancer patients by analyzing big data of the EHR (electronic health record) and EMR (electronic medical record). The phenotype classifier used NLP for feature extraction. The proposed system reported higher performance compared to the other techniques mentioned. Nevertheless, the results that the paper demonstrated were provided from simulated data; the authors will test real big data in their future work [29].
An NGS file preprocessing software named "FastProNGS" was introduced. It primarily integrates the quality control process with automatic adapter removal to preprocess NGS data files. The results have shown that the proposed software is faster than similar preprocessing tools.

Moreover, the authors reported that the introduced FastProNGS outperformed similar preprocessing tools in terms of processing time. Furthermore, the output can be formatted in plain-text, JSON, or HTML formats with user-friendly demonstration [30].

The serum exosome microRNAs of multiple sclerosis patients profiled using next-generation sequencing were studied as biomarkers for disease activity in response to fingolimod therapy. Disease activity was obtained through gadolinium-enhanced magnetic resonance imaging. Authors have used univariate/multivariate modeling and machine learning to obtain microRNA signatures. Moreover, the paper has reported individual microRNAs' modest power to predict disease status post fingolimod therapy (average 77%, range 65 to 91%) [22].

In this work, the same dataset used in [22] was utilized with the proposed model, and the obtained results outperformed the results reported in [22].

IV. MATERIALS AND METHODS

A. DEVELOPMENT ENVIRONMENT

The specifications of the development environment used to implement the introduced model are summarized in Table I.

TABLE I
DEVELOPMENT ENVIRONMENT

Specification	Value
Hardware accelerator	TPU
Available RAM	35.35GB
Disk Space	225.89GB
Development Environment	Python 3.6

B. DATASET

The dataset was available through the National Center for Biotechnology Information (NCBI) in November 2019 [23], consisting of NGS files of microRNA sequences of relapsing-remitting MS patients before fingolimod treatment and after six months of therapy. Dataset details are summarized in Table II. The data were balanced as 105 runs before the treatment and 110 total runs of patients after treatment, as illustrated in Figure 1.

TABLE II
DETAILS OF THE DATASET USED

Parameter	Value
Dataset source	NCBI
Dataset Provider	School of Biotechnology and Biomolecular Sciences, University of New South Wales
Dataset Registration Date	November 2019
Sequencing Instrument	Illumina HiSeq 2000
Strategy	ncRNA-Seq
Library Source	TRANSCRIPTOMIC
Sequencing Layout	Single
Source name	Serum_Exosome
Runs of untreated patients	105
Runs of treated patients	110
Number of total runs	215
Number of Patients	54
File type	SRA
Average FASTQ Files size	3 GB

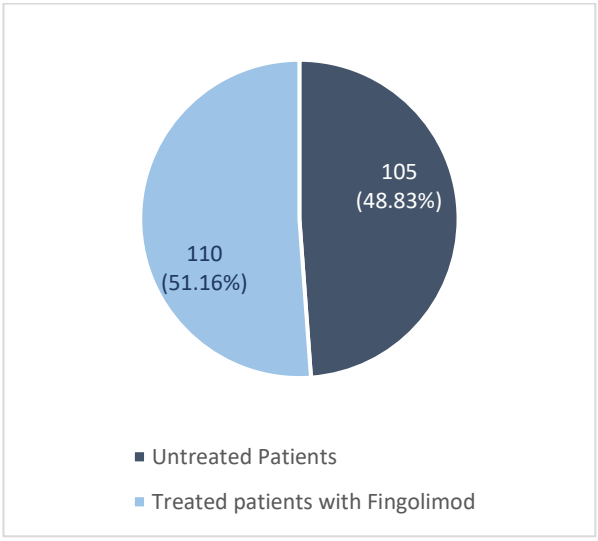

FIGURE 1. The Used MS patients Dataset balance- 51.16% of the files was of nontreated MS patients, and 48.83% of the files were of the patients after being treated for six months with Fingolimod

C. PROPOSED MODEL

In this work, a complete model for analyzing NGS data is proposed to obtain MS disease biomarkers by analyzing noncoding miRNA sequence files with nonbiological methods and combining text mining methods with predictive machine learning algorithms.

Several consecutive data preprocessing steps were applied to transform the given transcriptomic data into a text mining problem. Nevertheless, the data being analyzed are not in the English language and thus cannot use the ordinary feature extraction methods of English text. Hence, *KmerFIDF* was proposed as a feature extraction method, a hybridized K-mer counting method, and the term frequency inverse document frequency method.

As demonstrated in Figure 2, the proposed model is primarily divided into three main phases: preprocessing followed by the proposed feature extraction method and predictive machine learning algorithms. Each of these phases is explained in detail in the following sections.

D. SRA DATA DOWNLOAD

NCBI's studied dataset was directly downloaded to the cloud development platform using the *pysradb* library [24]. These procedures saved considerable time and helped to

avoid connectivity errors or data loss, incompletion, or interruption.

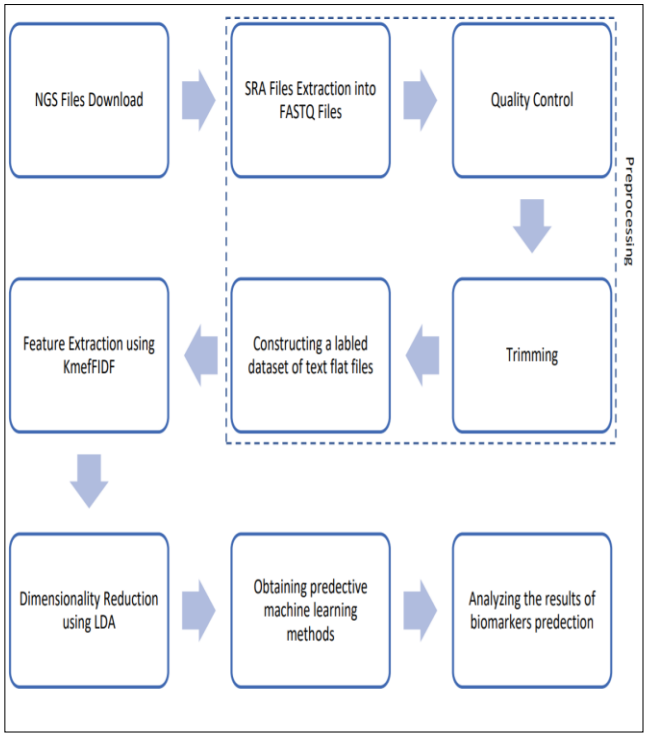

FIGURE 2. The proposed model for determining MS biomarkers by analyzing NGS files of microRNA data using nonbiological methods.

E. DATA PRE-PROCESSING

1) SRA FILES CONVERSION INTO FASTQ FILES

NCBI provides the NGS files in Sequence Read Archive (SRA) format. Consequently, it has to be converted into fastq files format. The "*SRA-Toolkit*" provided by NCBI was used for this file conversion process [25].

2) OBTAINING QUALITY CONTROL REPORTS

In this step, a quality check was applied to the converted files. The *FastQC* tool was used, which NCBI and literature recommend [26].

FastQC provides an inclusive view of the quality scores of the given input sequences, base composition statistics, length distribution in addition to other sequencing and preparation quality parameters such as trimming position and the contamination extent of data at specific positions of the file [27].

3) FILES TRIMMING

Based on this quality check, sequence trimming, filtering, and enhancement are performed. Trimming involves cropping sequence vectors with low quality and removing adapters from the sequences. As per NCBI and the literature, we used *Trimmomatic* for sequence trimming [28].

4) CONVERTING FASTQ FILES INTO FASTA FILES

Fastq files primarily contain nucleotide sequences and their quality scores; that is, they consist of a set of sequence reads. Each read has four lines. The first line begins with an "@" character followed by a sequence identifier. The second line contains the actual nucleotide sequence; the third line contains a '+' symbol followed optionally by the same sequence identifier. The fourth line encodes the quality values in ASCII for the sequence given in the second line.

On the other hand, Fasta files only contain nucleotide sequences in addition to their identifiers. Thus, this file conversion significantly shrinks the studied files' sizes.

The Biopython library was used to perform this file conversion accordingly [29][30].

5) DATASET FORMATION AND LABELING

In this step, the actual dataset construction takes place. That is, the transcriptome data are read from the produced fasta files. Hence, all sequence identifiers are removed, and then all the transcriptome sequences contained in one fasta file are scanned and concatenated together into one long transcriptome sequence. According to the NCBI-provided dataset metadata, a record label [treated/nontreated] is given to each produced sequence. Thus, in the following model steps, the transcriptome sequences shall be treated as plain throughput stream text [31].

F. FEATURE EXTRACTION USING THE KmerFIDF METHOD

1) OBTAINING K-MER COUNTS

K-mers are the unique subsequences of a k length nucleotide sequence. Table III demonstrates the K-mer method's notion. As shown, a sequence of a length L can produce L − k + 1 k-mers [32].

Hence, the K-mer method was applied to segment the transcriptome producing the "K-mer matrix" of M Sequences × 4^k *(i.e., given the four proteins A, C, T, and G).* Following this k-mer segmentation, counting vectors were obtained accordingly for each k-*mer sequence* in each *fasta* file of the studied dataset.

TABLE III
POSSIBLE PRODUCED K-MERS (TERMS) GIVEN A SIMPLE NUCLEOTIDE SEQUENCE (CTGAACTT) USING THE K-MER METHOD

K	k-mers	possible k-mers (L-k+1)
1	C,T,G,A,A,C,T,G	8
2	CT,TG,GA,AA,AC,CT,TT	7
3	CTG,TGA,GAA,AAC,ACT,CTT	6
4	CTGA,TGAA,GAAC,AACT,ACTT	5
5	CTGAA,TGAAC,GAACT,AACTT	4
6	CTGAAC,TGAACT,GAACTT	3
7	CTGAACT,TGAACCTT	2
8	CTGAACTT	1

2) OBTAINING THE PROPOSED KmerFIDF FOR FEATURE EXTRACTION

In data mining, Term Frequency Inverse Document Frequency (TFIDF) is a numerical statistic method to imply a given term's weight within a given document [33].

In this work, this method was combined along with the K-mer counting method for transcriptomic feature extraction. *KmerFIDF* is demonstrated in detail as follows:

The inverse document frequency *IDF* and *TFIDF* are defined by equations (1) and (2), respectively. where *t* is the term to be evaluated, *d* is the document, *N* is the number of documents in the entire corpus *D*, and *tf(t,d)* is the frequency of a term *t* in document *d* [34].

$$idf(t, D) = log \frac{N}{|\{d \epsilon D : t \epsilon d\}|} \quad (1)$$

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \quad (2)$$

The preceding processing step segmented the transcriptomic sequence into *kmers*, and kmer frequency vectors were attained for each k-mer sequence in each fasta file in the dataset.

Our method *KmerFIDF* defined by equation (3) involves using these *KmerCount* vectors as a substitutional parameter in the original TFIDF equation. This new equation has retained the sequence ordering and respected the possible combinations of the produced *kmers*. *idf(t, D)* is obtained given the *KmerC*ount matrix to determine the features' (i.e., *Kmers*) weights.

$$KmerFIDF(t, d, D) = KmerCount(t, d) \cdot idf(t, D) \quad (3)$$

G. DIMENSIONALITY REDUCTION

Ordinarily, the preceding step obtains a sparse matrix. Hence, dimensionality reduction has to be attained to determine the features with higher weights.

LDA can efficiently perform dimensionality reduction of supervised classification problems since LDA leverages the maximization separability among the classes which gives it superiority with text mining problems as per literature [35][36]. In addition, LDA primarily projects the data in a new linear feature space. Consequently, the classifier shall reach high accuracy if the data are linearly separable. Unlike other dimensionality reduction algorithms, such as principal component analysis (PCA), which maximizes the variance of the data within a class [37].

In the proposed model, the linear discriminant analysis (LDA) algorithm is used for dimensionality reduction. Hence, the d-dimensional mean vector is computed for each class. Then, between-class scatter matrix S_w and the in-class scatter matrix S_B are constructed accordingly. S_w is given by equations (4) and (5). where c is the number of classes, x is the feature vector of each sequence file (*dataset row*), m_i is the mean vector of class $i \in c$ and *D* is the studied dataset. At the same time, S_B is given by equation (6), where N is the number of samples and *m* is the overall computed mean, including all samples from all classes. Afterward, the eigenvalue is calculated accordingly using equation (7). The eigenvalues are then sorted, and the lowest values are eliminated [38].

$$S_w = \sum_{i=1}^{c} S_i \quad (4)$$

$$S_i = \sum_{x \in D_i}^{c} (x - m_i)(x - m_i)^T \quad (5)$$

$$S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T \quad (6)$$

$$\text{Eigenvalue} = S_w^{-1} S_B \quad (7)$$

H. SUPERVISED MACHINE LEARNING PREDICTIVE METHODS

As per the literature, three machine learning algorithms were applied: SVM, logistic regression (LR), and random forest (RF) [39]. A *sklearn* library was used for the implementation of these three methods. Table IV summarizes the significant parameters that were given a value other than their default values. Additionally, Table V summarizes the essential libraries and tools used in the proposed model implementation.

To ensure model robustness and overcome the possibility of overfitting and selection bias in the proposed model, cross-validation was applied with each of the implemented predictive methods with a folding value of 10. Additionally, the dataset was split into 30% for testing and 70% for training.

TABLE IV
THE SIGNIFICANT PARAMETERS THAT WERE DEFINED IN THE IMPLEMENTED PREDICTIVE MODEL

Predictive Method	Parameter	Value
SVM, LR, & RF	Cross-validation Folding	10
SVM	Kernel	Linear
RF	Number of Estimators	500
RF	Information gain method	Gini
Model Preparation	Test Size	0.3

TABLE V
ESSENTIAL TOOLS AND LIBRARIES USED

Implementation Step	Tool/Library Used
Data Download	*Pysradb*
Files Quality Control	*FastQC*
Files Trimming	*Trimmomatic*
File Fastq files conversion	*Biopython*
Predictive models implementation	*Sklearn*

V. RESULTS AND DISCUSSION

Upon implementing the introduced model, three sets of experiments were conducted. The first was for parameters tuning. The second set of experiments was undertaken to evaluate the proposed model in terms of detection accuracy and execution time. Moreover, the third set of experiments

was conducted to evaluate the proposed model against state-of-the-art methods.

The *K* parameter of *KmerFIDF* was tuned to $k \in \{2,3,4,5,6\}$; obtaining the value of $K>6$ was not feasible due to processing constraints.

Figure 3 illustrates the obtained average accuracy scores (*over 10 runs for each experiment*) denoted by applying each of the three predictive models with each value of $k \in \{2,3,4,5,6\}$. As shown, the highest accuracy scores were obtained when k=5. RF (97.0), while SVM and LR were 94.0. The demonstrated k-parameter tuning results denote that the prediction accuracy is proportionally increased with K's value until *k*=5 and k=6. This relationship can be justified since the higher the value of *k* is, the more contextual information is in each feature.

The second parameter tuning experiment was to determine the "*number of estimators*" parameter, which denotes the number of trees in the RF model. As illustrated in Figure 4, experiments of 200-700 estimators were run. The peak point of average accuracy was obtained by 500 estimators, with no further accuracy enhancement on increasing the number of estimators above 500.

Furthermore, Figure 5 summarizes the obtained results of the second set of experiments that evaluate the model performance. The reported values of sensitivity, specificity, F1-score, and average accuracy scores with RF were 96.4, 96.47, 95.6, and 97%, respectively; 92.89, 92.78, 93.2, and 94% with SVM, respectively; and 91.95, 92.2, 93.1, and 94% with LR, respectively.

As illustrated, RF has reported the highest accuracy due to the random feature selection over the numerous features set being analyzed. The trees are more independent, yielding better predictive performance because of the higher variance-bias trade-offs where each tree learns only from a subset of features.

On the other hand, SVM primarily works to define a hyperplane over a given dataset. Consequently, this margin maximization is more challenging on a high-dimensional dataset. Similarly, LR assumes a linear relationship among data points within the studied dataset. There is no severe multicollinearity among the explanatory variables. Consequently, RF has reported the highest specificity, sensitivity, and F1-score.

In terms of execution time, no significant difference was reported across the experimented predictive models. The execution time ranged from 89 to 90.5 minutes. Figure 6 summarizes the average computational times.

The third set of experiments was conducted to compare the obtained model accuracy with the results of two other studies that aimed to detect MS disease. The first literature work studied the exact dataset used to evaluate our model. It used conventional univariate/multivariate modeling for feature extraction. Univariate analysis involves analyzing a single variable, and multivariate analysis uses two or more dependent variables and multiple independent variables. This literature study also applied RF as a predictive model. Figure 7 shows a comparative chart of the average accuracy scores reported by our proposed model using SVM, RF, and LR as 94, 97, and 94, respectively. In comparison, the average accuracy score and highest accuracy reported by referenced work using RF were 77.0 and 91.0, respectively [22].

As illustrated, analyzing microRNA data using the proposed model has outperformed the referenced work[22]. This accuracy difference can be clearly explained, as our proposed model has applied feature engineering among the entire dataset. On the other hand, the referenced work used conventional univariate/multivariate modeling. Furthermore, the obtained experimental results indicate that biomarkers of multiple sclerosis disease can be obtained from analyzing microRNA data. This finding also confirms the impact of fingolimod treatment on MS biomarkers.

Finally, the reported results of the proposed model were compared to the results obtained from a second study. This literature work studied MS disease detection by analyzing EEG signal data of MS patients with the K-NN classifier algorithm and reported overall accuracy, sensitivity, and specificity of 80%, 72.7%, and 88.9%, respectively [40]. This result implies that the introduced model outperformed this literature study. Two key aspects can justify this. The first is that the studied transcriptomic data have more indicative features of the disease biomarkers. The second is the robustness of the proposed model in the detection of MS disease biomarkers. Figure 8 summarizes the reported accuracy, sensitivity, and specificity scores reported by the introduced and referenced studies.

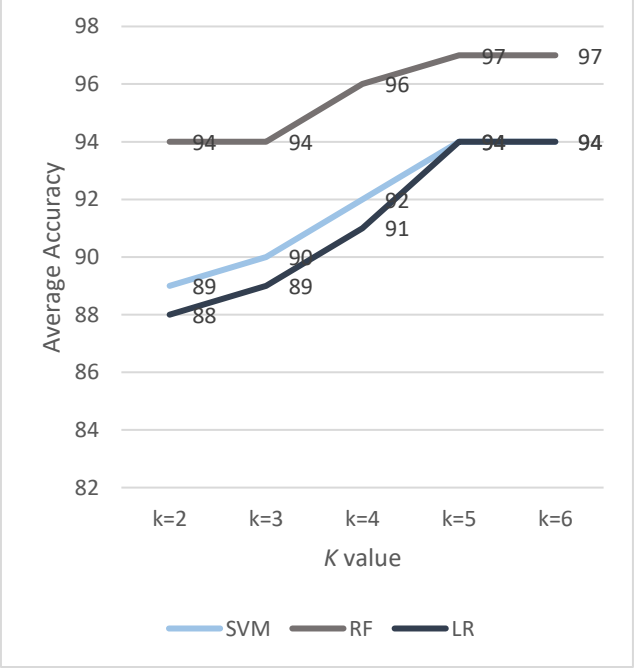

FIGURE 3: The average accuracy scores denoted by the parameter tuning experiments of K value

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3109069, IEEE Access

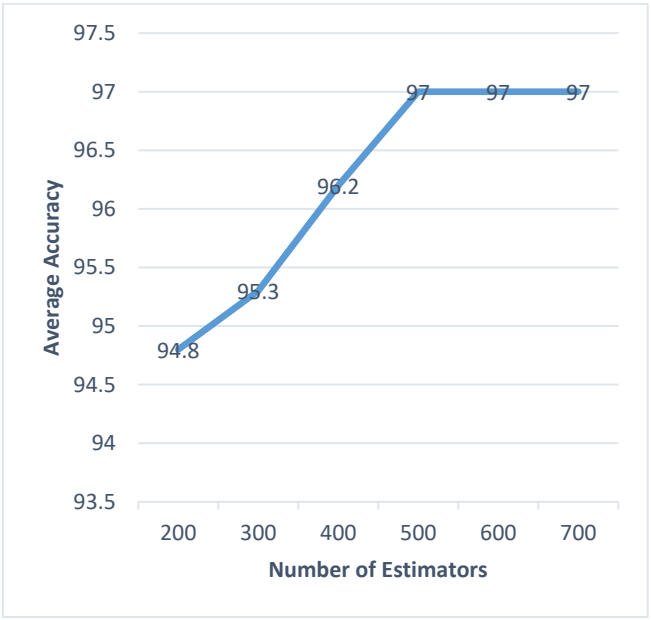

FIGURE 4: Tuning the "number of estimators" parameter of the random forest while keeping the K values of KmerFIDF = 6

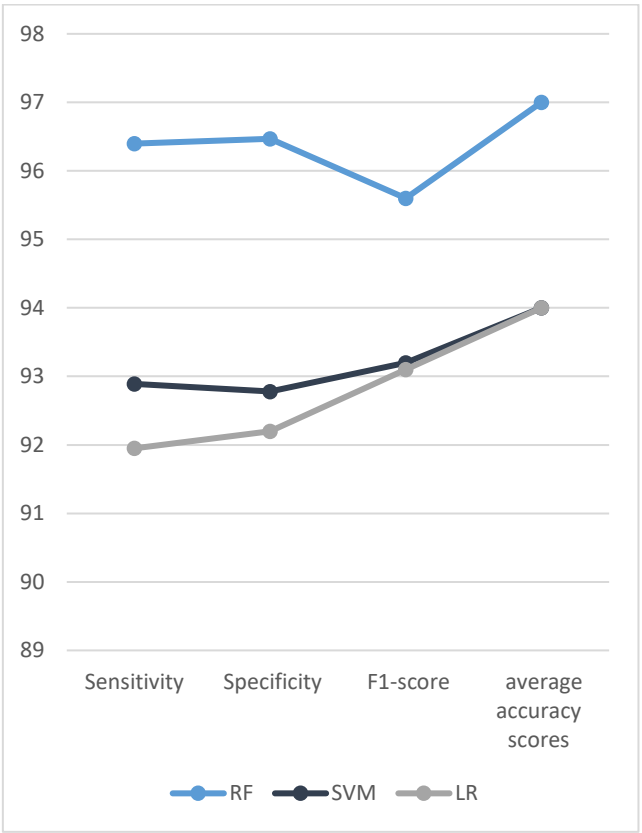

FIGURE 5. Sensitivity, Specificity, F1-Score, and Average Accuracy Scores of RF, SVM, and LR

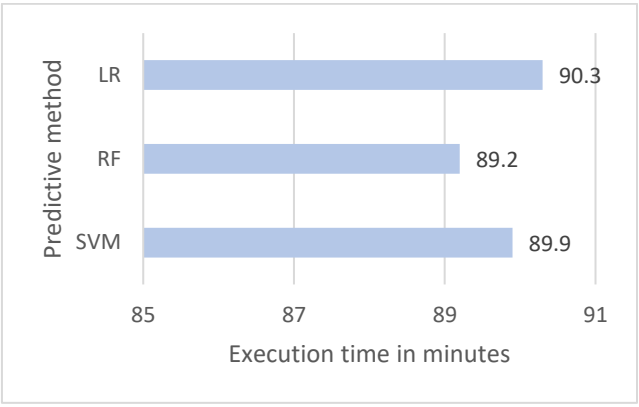

FIGURE 6: The obtained execution time (in minutes) of each predictive method (LR, RF, and SVM) used with the proposed model.

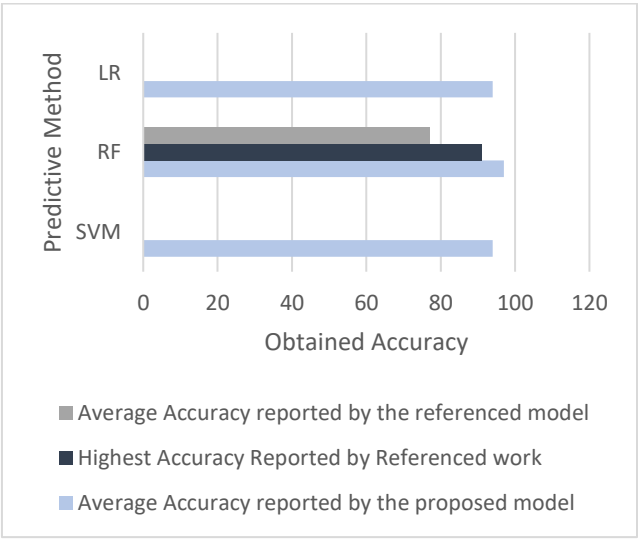

FIGURE7 . Comparing the results reported by the proposed model using (SVM, RF and LR) methods with a reference method (using the RF method) reported in [22].

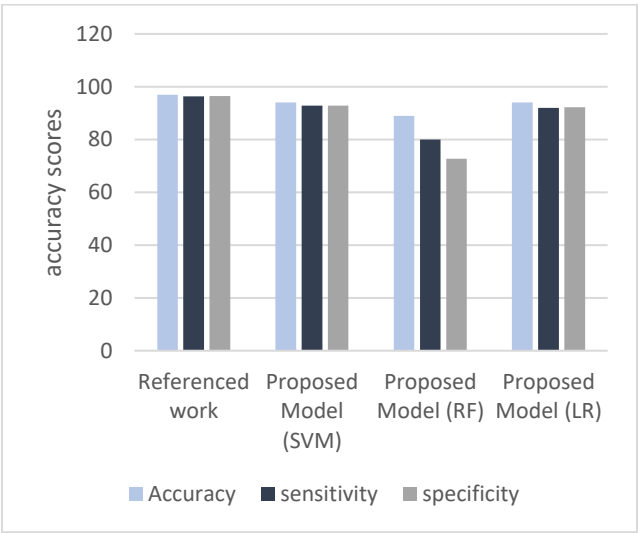

FIGURE 8. Reported (overall accuracy, sensitivity, specificity) of referenced work [40] **and the proposed model.**

VI. CONCLUSIONS AND FUTURE WORK

This work introduced a detailed model for multiple sclerosis disease biomarker detection by analyzing transcriptomic microRNA data through transforming the phenotype classification problem into a text mining problem. Experimental work was applied to a transcriptomic dataset of multiple sclerosis patients before fingolimod treatment and six months after treatment. The highest reported sensitivity, specificity, F1-score, and average accuracy scores were 96.4, 96.47, 95.6, and 97%, respectively, indicating auspicious results in disease biomarker detection from transcriptomic data.

Moreover, this work introduced the *KmerFIDF* method as a novel feature extraction method and has applied comparative experiments with two literature works. The implications of this model indicate that the proposed model outperformed the first literature method over the same dataset and used the same random forest as a predictive method. The introduced model reported an average accuracy on the random forest algorithm of 97%, while the literature model reported an average accuracy score of 77% and 91% as the highest accuracy. Furthermore, the introduced experimental results confirm that fingolimod treatment decreases disease progression given the considerable classification accuracy.

Finally, the experimental results also denote that the proposed model outperforms the traditional EEG analysis for MS diagnosis.

In our future work, further experimental work will be performed on this model using other coding and noncoding transcriptomic datasets of MS disease. Our proposed model will also be tested against transcriptomic datasets of other diseases to study the model's robustness. Furthermore, it enables the model to process more extensive sizes of transcriptomic files. Data preprocessing enhancements, including other dimensionality reduction methods, shall also be examined and considered.

REFERENCES

[1] B. Song and Y. Shu, "Cell vibron polariton resonantly self-confined in the myelin sheath of nerve," *Nano Res.*, vol. 13, no. 1, pp. 38–44, Jan. 2020, doi: 10.1007/s12274-019-2568-4.

[2] M. M. A. El Hamid, N. M. Ali, M. N. Saad, M. S. Mabrouk, and O. G. Shaker, "Multiple sclerosis: an associated single-nucleotide polymorphism study on Egyptian population," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 9, no. 1, p. 48, Dec. 2020, doi: 10.1007/s13721-020-00255-6.

[3] C. Y. Xia *et al.*, "Connexins in oligodendrocytes and astrocytes: Possible factors for demyelination in multiple sclerosis," *Neurochemistry International*, vol. 136. Elsevier Ltd, p. 104731, Jun. 01, 2020, doi: 10.1016/j.neuint.2020.104731.

[4] N. M. Ali, M. Shaheen, M. S. Mabrouk, and M. A. Aborezka, "MACHINE LEARNING IN EARLY GENETIC DETECTION OF MULTIPLE SCLEROSIS DISEASE: A SURVEY," *Int. J. Comput. Sci. Inf. Technol.*, vol. 12, no. 5, 2020, doi: 10.5121/ijcsit.2020.12501.

[5] M. L. Elkjaer *et al.*, "Molecular signature of different lesion types in the brain white matter of patients with progressive multiple sclerosis," *Acta Neuropathol. Commun.*, vol. 7, no. 1, Dec. 2019, doi: 10.1186/s40478-019-0855-7.

[6] N. Zhang, G. Hu, T. G. Myers, and P. R. Williamson, "Protocols for the Analysis of microRNA Expression, Biogenesis, and Function in Immune Cells," *Curr. Protoc. Immunol.*, 2019, doi: 10.1002/cpim.78.

[7] J. P. Yin Lee *et al.*, "Gene expression profiling of giant fibroadenomas of the breast," *Surg. Oncol.*, vol. 37, p. 101536, Jun. 2021, doi: 10.1016/j.suronc.2021.101536.

[8] D. van den Heuvel, Y. van der Weegen, D. E. C. Boer, T. Ogi, and M. S. Luijsterburg, "Transcription-Coupled DNA Repair: From Mechanism to Human Disorder," *Trends in Cell Biology*, vol. 31, no. 5. Elsevier Ltd, pp. 359–371, May 01, 2021, doi: 10.1016/j.tcb.2021.02.007.

[9] D. A. Garcia *et al.*, "An intrinsically disordered region-mediated confinement state contributes to the dynamics and function of transcription factors," *Mol. Cell*, vol. 81, no. 7, pp. 1484-1498.e6, Apr. 2021, doi: 10.1016/j.molcel.2021.01.013.

[10] X. Yu *et al.*, "Messenger RNA 5′ NAD+ Capping Is a Dynamic Regulatory Epitranscriptome Mark That Is Required for Proper Response to Abscisic Acid in Arabidopsis," *Dev. Cell*, vol. 56, no. 1, pp. 125-140.e6, Jan. 2021, doi: 10.1016/j.devcel.2020.11.009.

[11] N. Li *et al.*, "Cleavage and polyadenylation-specific factor 3 induces cell cycle arrest via PI3K/Akt/GSK-3β signaling pathways and predicts a negative prognosis in hepatocellular carcinoma," *Biomark. Med.*, vol. 15, no. 5, pp. 347–358, Apr. 2021, doi: 10.2217/bmm-2021-0039.

[12] A. E. Deveaux *et al.*, "RNA splicing and aggregate gene expression differences in lung squamous cell carcinoma between patients of West African and European ancestry," *Lung Cancer*, vol. 153, pp. 90–98, Mar. 2021, doi: 10.1016/j.lungcan.2021.01.015.

[13] P. H. Wu and P. D. Zamore, "To Degrade a MicroRNA, Destroy Its Argonaute Protein," *Mol. Cell*, vol. 81, no. 2, pp. 223–225, Jan. 2021, doi: 10.1016/j.molcel.2020.12.043.

[14] J. Huang *et al.*, "Tissue-specific reprogramming of host tRNA transcriptome by the microbiome," *Genome Res.*, p. gr.272153.120, Apr. 2021, doi: 10.1101/gr.272153.120.

[15] U. Naseem, M. Khushi, S. K. Khan, K. Shaukat, and M. A. Moni, "A comparative analysis of active learning for biomedical text mining," *Appl. Syst. Innov.*, vol. 4, no. 1, p. 23, Mar. 2021, doi: 10.3390/asi4010023.

[16] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, "The application of text mining methods in innovation research: current state, evolution patterns, and development priorities," in *R and D Management*, Jun. 2020, vol. 50, no. 3, pp. 329–351, doi: 10.1111/radm.12408.

[17] J. Rashid *et al.*, "Topic Modeling Technique for Text Mining over Biomedical Text Corpora through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering," *IEEE Access*, vol. 7, pp. 146070–146080, 2019, doi: 10.1109/ACCESS.2019.2944973.

[18] R. E. Nelson, J. Butler, J. Lafleur, K. Knippenberg, A. W. C. Kamauu, and S. L. Duvall, "Determining Multiple Sclerosis Phenotype from Electronic Medical Records," 2016. Accessed: Jan. 14, 2020. [Online]. Available: www.jmcp.org.

[19] C. A. Turner *et al.*, "Word2Vec inversion and traditional text classifiers for phenotyping lupus," *BMC Med. Inform. Decis. Mak.*, vol. 17, p. 126, 2017, doi: 10.1186/s12911-017-0518-1.

[20] H. S. Chase, L. R. Mitrani, G. G. Lu, and D. J. Fulgieri, "Early recognition of multiple sclerosis using natural language processing of the electronic health record," *BMC Med. Inform. Decis. Mak.*, 2017, doi: 10.1186/s12911-017-0418-4.

[21] E. Aun, A. Brauer, V. Kisand, T. Tenson, and M. Remm, "A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria," *PLoS Comput. Biol.*, 2018, doi: 10.1371/journal.pcbi.1006434.

[22] S. Ebrahimkhani *et al.*, "Serum Exosome MicroRNAs Predict Multiple Sclerosis Disease Activity after Fingolimod Treatment," *Mol. Neurobiol.*, 2020, doi: 10.1007/s12035-019-01792-6.

[23] "Serum exosome microRNAs predict multiple sclerosis... (ID 588268) - BioProject - NCBI." https://www.ncbi.nlm.nih.gov/bioproject/PRJNA588268

(accessed Nov. 15, 2020).

[24] S. Choudhary, "pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive," *F1000Research*, vol. 8, p. 532, Apr. 2019, doi: 10.12688/f1000research.18676.1.

[25] S. Utturkar, A. Dassanayake, S. Nagaraju, and S. D. Brown, "Bacterial differential expression analysis methods," in *Methods in Molecular Biology*, vol. 2096, Humana Press Inc., 2020, pp. 89–112.

[26] F. Contaldi, E. Cappetta, and S. Esposito, "Practical Workflow from High-Throughput Genotyping to Genomic Estimated Breeding Values (GEBVs)," in *Methods in Molecular Biology*, vol. 2264, Humana Press Inc., 2021, pp. 119–135.

[27] P. Vats, A. M. Chinnaiyan, and C. Kumar-Sinha, "Case Study: Systematic Detection and Prioritization of Gene Fusions in Cancer by RNA-Seq: A DIY Toolkit," in *Methods in Molecular Biology*, vol. 2079, Humana Press Inc., 2020, pp. 69–79.

[28] J. L. Sevigny, J. L. Norenburg, and F. Leasi, "A Bioinformatics Tutorial for Comparative Development Genomics in Diverse Meiofauna," in *Methods in Molecular Biology*, vol. 2219, Humana Press Inc., 2021, pp. 289–305.

[29] T. C. Cornish, L. J. Kricka, and J. Y. Park, "A Biopython-based method for comprehensively searching for eponyms in Pubmed," *MethodsX*, vol. 8, p. 101264, Feb. 2021, doi: 10.1016/j.mex.2021.101264.

[30] S. M. Ireland and A. C. R. Martin, "atomium—a Python structure parser," *Bioinformatics*, vol. 36, no. 9, pp. 2750–2754, May 2020, doi: 10.1093/bioinformatics/btaa072.

[31] B. Saremi, M. Kohls, P. Liebig, U. Siebert, and K. Jung, "Measuring reproducibility of virus metagenomics analyses using bootstrap samples from FASTQ-files," *Bioinformatics*, Nov. 2020, doi: 10.1093/bioinformatics/btaa926.

[32] J. Ge, J. Meng, N. Guo, Y. Wei, P. Balaji, and S. Feng, "Counting Kmers for Biological Sequences at Large Scale," *Interdiscip. Sci. Comput. Life Sci.*, vol. 12, no. 1, pp. 99–108, Mar. 2020, doi: 10.1007/s12539-019-00348-5.

[33] A. P. Pimpalkar and R. J. Retna Raj, "Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features," *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.*, vol. 9, no. 2, pp. 49–68, Jun. 2020, doi: 10.14201/adcaij2020924968.

[34] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *Int. J. Digit. Libr.*, vol. 17, no. 4, pp. 305–338, Nov. 2016, doi: 10.1007/s00799-015-0156-0.

[35] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Comput. Sci. Rev.*, vol. 40, p. 100378, May 2021, doi: 10.1016/J.COSREV.2021.100378.

[36] D. Wu, R. Yang, and · Chao Shen, "Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm," doi: 10.1007/s10844-020-00597-7.

[37] S. K. Das, T. S. Bhattacharya, M. Ghosh, and J. Chowdhury, "Probing blood plasma samples for the detection of diabetes using SERS aided by PCA and LDA multivariate data analyses," *New J. Chem.*, vol. 45, no. 5, pp. 2670–2682, Feb. 2021, doi: 10.1039/d0nj04508j.

[38] R. Rani and D. K. Lobiyal, "An extractive text summarization approach using tagged-LDA based topic modeling," *Multimed. Tools Appl.*, vol. 80, no. 3, pp. 3275–3305, Jan. 2021, doi: 10.1007/s11042-020-09549-3.

[39] N. Mohamed Ali, M. M. A. El Hamid, and A. Youssif, "SENTIMENT ANALYSIS FOR MOVIES REVIEWS DATASET USING DEEP LEARNING MODELS," *Int. J. Data Min. Knowl. Manag. Process*, 2019, doi: 10.5121/ijdkp.2019.9302.

[40] B. K. Karaca, M. F. Akşahin, and R. Öcal, "Detection of multiple sclerosis from photic stimulation EEG signals," *Biomed. Signal Process. Control*, vol. 67, p. 102571, May 2021, doi: 10.1016/j.bspc.2021.102571.

Nehal M. Ali was born in New Jersey, NJ, USA. She received her B.Sc. in Computer Science from Helow University, Cairo, Egypt, in 2006, and her M.Sc. degree in computer science from the Arab Academy for Science Technology and Maritime Transport, Egypt, in 2013. She is currently working on her Ph.D. degree in bioinformatics. In addition to being a software engineer and a data scientist for more than 14 years in a multinational corporation, she also has several publications in bioinformatics, machine learning, and text mining fields. She is also a reviewer for multiple scientific journals and has won several recognitions in national and international competitions.

Mohamed Shaheen received a B.Sc. degree in computer science from Alexandria University, Egypt, 1991, an M.Sc. degree in computer engineering from Arab Academy for Science, Technology and Maritime Transport (AASTMT), Alexandria, Egypt, 1998, as well as an M.Sc. and the Ph.D. degrees in computer engineering from University of Louisiana at Lafayette, Louisiana, USA, in 2000 and 2003, respectively.
He is currently a Professor at the College of Computing and Information Technology, AASTMT, since 2010. From 2003-2010, he has been serving as Graduate Faculty at the University of Louisiana at Lafayette, USA. From 1999 to 2003, he served as an Adjunct Faculty member in the computer science department at the University of Louisiana at Lafayette, USA. His research interests include Software Engineering, Wireless Sensor Networks, and Artificial Intelligence. His research is supported and funded by federal agencies including the DOE, NSF, the Louisiana Governor ITI, the Qatar National Research Fund ("QNRF"), and ITIDA, Egypt.

Mai S. Mabrouk received her B.Sc., M.Sc., and Ph.D. degrees from the Biomedical Engineering Department at Cairo University in 2000, 2004, and 2008. She is currently an Associate professor of and department head of Biomedical Engineering at Misr University for Science and Technology. Her biography was selected to appear in Marquis Who's Who in the World in 2012. She has served as a technical reviewer and editorial board member for several international journals and conferences throughout her career. She has published over 90 peer-reviewed journal and conference articles in medical imaging processing, Bioinformatics, and the human-computer interface.

Mohamed a. Aborizka received his BS degree from MTC, Cairo, Egypt, in 1989, an MS degree from MTC, Cairo, Egypt, in 1995, and Ph.D. degree from the University of Alabama Huntsville, Alabama, United States, in 2002. He was a Lecturer in MTC from 2002-2004. He was Chairman of the eCommerce Department, Faculty of Management and Information Technology, Arab Academy for Science Technology & Maritime Transport (AASTMT), Cairo, Egypt 2004-2007. He was also the associate dean of the Faculty of Management and Information Technology, AASTMT, from 2007-2011. From 2011 to 2017, he was a Dean of Center of Excellence, AASTMT. From 2017 till currently, he is a Dean of College of Computing and Information Technology (AASTMT). His research interests include Big Data Analytics, Intelligent Systems, E-learning, Bioinformatics, Distributed Systems, and Cloud Computing.