An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers

USMAN NASEEM[1], **JUNAID RASHID**[2], **LIAQAT ALI**[3], **JUNGEUN KIM**[2,4],
QAZI EMAD UL HAQ[5], **MAZHAR JAVED AWAN**[6], **AND MUHAMMAD IMRAN**[7]

[1]Department of Information Technology, Sydney International School of Technology and Commerce, Sydney, NSW 2000, Australia
[2]Department of Computer Science and Engineering, Kongju National University, Cheonan 31080, South Korea
[3]Department of Electrical Engineering, University of Science and Technology Bannu, Bannu 28100, Pakistan
[4]Department of Software, Kongju National University, Cheonan 31080, South Korea
[5]Center of Excellence in Cybercrime and Digital Forensics (CoECDF), Naif Arab University for Security Sciences (NAUSS), Riyadh 14812, Saudi Arabia
[6]Department of Software Engineering, University of Management and Technology, Lahore 54770, Pakistan
[7]School of Engineering, Information Technology and Physical Sciences, Federation University, Brisbane, QLD 4000, Australia

Corresponding authors: Junaid Rashid (junaidrashid062@gmail.com) and Jungeun Kim (jekim@kongju.ac.kr)

ABSTRACT Breast cancer (BC) is the second most prevalent type of cancer among women leading to death, and its rate of mortality is very high. Its effects will be reduced if diagnosed early. BC's early detection will greatly boost the prognosis and likelihood of recovery, as it may encourage prompt surgical care for patients. It is therefore vital to have a system enabling the healthcare industry to detect breast cancer quickly and accurately. Machine learning (ML) is widely used in breast cancer (BC) pattern classification due to its advantages in modelling a critical feature detection from complex BC datasets. In this paper, we propose a system for automatic detection of BC diagnosis and prognosis using ensemble of classifiers. First, we review various machine learning (ML) algorithms and ensemble of different ML algorithms. We present an overview of ML algorithms including ANN, and ensemble of different classifiers for automatic BC diagnosis and prognosis detection. We also present and compare various ensemble models and other variants of tested ML based models with and without up-sampling technique on two benchmark datasets. We also studied the effects of using balanced class weight on prognosis dataset and compared its performance with others. The results showed that the ensemble method outperformed other state-of-the-art methods and achieved 98.83% accuracy. Because of high performance, the proposed system is of great importance to the medical industry and relevant research community. The comparison shows that the proposed method outperformed other state-of-the-art methods.

INDEX TERMS Healthcare system, machine learning, breast cancer, ensemble learning, cancer diagnoses.

I. INTRODUCTION

Breast cancer is one of the most dangerous and prevalent cancers among women, causing the deaths of large numbers of women worldwide. Breast cancer accounts for 8.4% of diagnosed cancers and 6.6% of cancer-related deaths worldwide, according to a World Health Organization (WHO) report [1]. Breast cancer accounted for 15.9% of all reported cancers

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed Farouk.

among Saudi citizens and 28.7% of all reported cancers among women of all ages, according to the Saudi Health Council [2]. Breast cancer is more common in women with dense breasts, and there is a relationship between density and age, with younger women having denser breasts than older women [3]. The American College of Radiology developed the Breast Imaging Data and Reporting System (BI-RADS). Table 1 presents the four BI-RADS assessment categories. Despite recent developments in computer vision, screening mammography is still read and interpreted manually under

TABLE 1. BI-RADS assessment categories [4].

BI-RADS CATEGORY	Breast density	Risk of CANCER (%)
BI-RADSI	Almost entirely fatty	(0-25%)
BI-RADSII	Some fibro glandular tissue	(26-50%)
BI-RADSIII	Heterogeneously dense	(51-75%)
BI-RADSIV	Extremely dense	(76-100%)

the supervision of a radiologist. However, the enormous number of screening images is difficult for radiologists to handle accurately.

Recently, medical imaging researchers have used state-of-the-art techniques to solve problems with breast cancer analysis [4]. Several research has been carried out on the automatic identification of breast cancer. According to the World Health Organization, BC is the most frequently diagnosed cancer in women, accounting for approximately one in four newly diagnosed cancer cases. According to the World Cancer Research Fund International, only 1.7 million new cases were reported in 2012 (WCRF). Despite its high incidence and even the absence of early signs [5], early identification of BC can considerably improve the chances of survival. According to the WCRF, patients identified with stage I/II BC have a five-year survival rate of 80-90 percent, while patients diagnosed with stage III/IV BC have a survival rate of only 24 percent. As a result, it is obvious that the proper classification of benign tumors is essential to encourage patients to seek appropriate therapy and obtain a better prognosis. As a result, considerable research in the diagnosis of BC focuses on accurately identifying individuals as malignant or benign. Many machine learning (ML) algorithms and neural network (NN) approaches have been used in the BC Wisconsin diagnostic and prognostic dataset. For the classification challenge, researchers have presented a large number of ML techniques in previous articles. We fully describe the various classification algorithms used to classify BC in this investigation. We mainly focus on artificial neural network (ANN) methods based on deep learning (DL), as well as support vector machines (SVM) based on traditional machine learning (ML) and k-nearest neighbor (KNN) and decision tree (DT) algorithms.

BC data has lots of varieties, open-source and private included. In the 1990s, a Wisconsin hospital collected three datasets, one of which was WBCD, which has strong representation and comparability due to having many algorithms applied to it. 94.36% to 99.90% is the range within which the classification accuracies of different algorithms are achieved. Developing different ML algorithms that are better improved is still needed to obtain alternative answers to complex BC data in the real world or other medical data, even though very high accuracies are achievable with a significant number of already existing algorithms. Though not the only criterion, classification accuracy is intuitive and especially important. The need for new algorithms to be developed to improve the

existing techniques, considering that they still have drawbacks despite having their specific advantages. Accuracy in physicians' decision-making is greatly aided by these algorithms once used to build healthcare systems that can provide second opinions. Open-source benchmark databases beginning with ANNs, then SVM, DTs, and k-NNs are analyzed by the ML algorithms highlighted in the coming subsections. In this paper, our main contributions are:

- We presented an ensemble of machine learning-based methods for breast cancer diagnosis and prognosis using an ensemble of machine learning classifiers.
- We presented a comprehensive comparison of the performance of various machine learning and ensemble machine learning-based classifiers.
- We evaluated different sampling methods to address the class imbalance issue in our datasets.
- We demonstrated that the proposed method outperforms various state-of-the-art methods for the detection of breast cancer.
- Analysis with and without sampling techniques is performed.

The rest of this paper is organized as follows: the background on breast cancer and computer-assisted diagnostics is presented in Section 2. The literature review is presented in Section 3. The proposed methodology is described in Section 4. The results and analysis of the experiments are discussed in Section 5. In Section 6, we conclude our work.

II. BACKGROUND
A. BREAST CANCER
Breast cancer is a disease in which malignant (cancer) cells grow in the breast tissues. A tumor is a mass of diseased tissue. There are two types of breast tumors: non-cancerous, "benign," and cancerous, or "malignant." Cancer starts in the cells that are the basic building blocks in the breast or other body parts that make up tissue. Occasionally the way of cell outgrowth goes fault, and new cells form or old or damaged cells would not die as they do when the body does not need them [7]. Any new breast, lump, or breast changes should then be monitored by a health care professional experienced in the diagnosis of breast disease that is commonly a sign of breast cancer [8]. No treatment has yet been discovered for cancer tumors. However, early detection of breast cancer is crucial to minimize the number of cancer deaths and enhance patients' quality of life. On a mass can be 'circumscribed', 'Micro-lobulated', 'Obscured', 'Indistinct', or 'Speculated' as shown in Figure 2.

The most widespread and recognized factors increasing the risk of breast cancer are female sex, age, genetics, and having dense breasts. Breast density is a measure used to depict the extent of the distinctive tissues that form a woman's breasts and how the breasts look on a mammogram. Breast cancer is more common in women with dense breasts, and there is a relationship between densities and women's age, with younger women having denser breasts than older women [3].
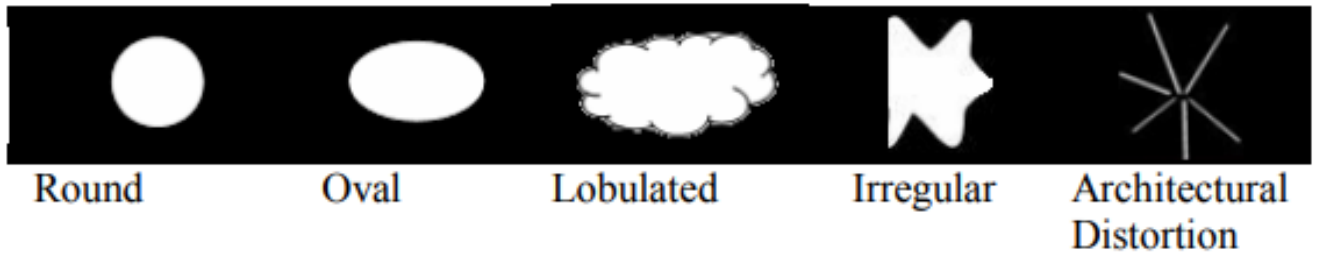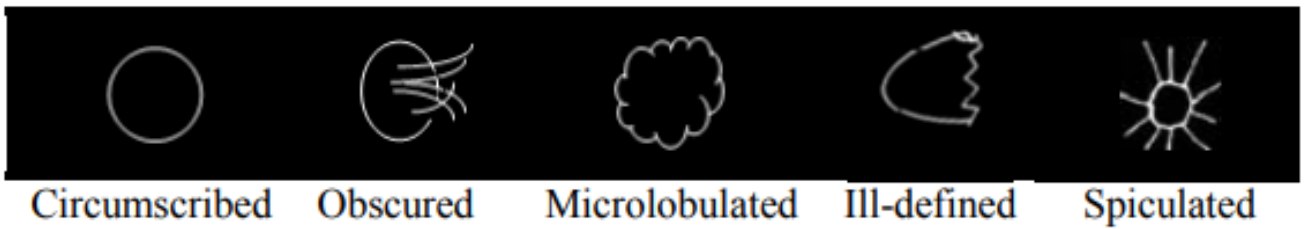
FIGURE 1. Shape of breast mass [6].

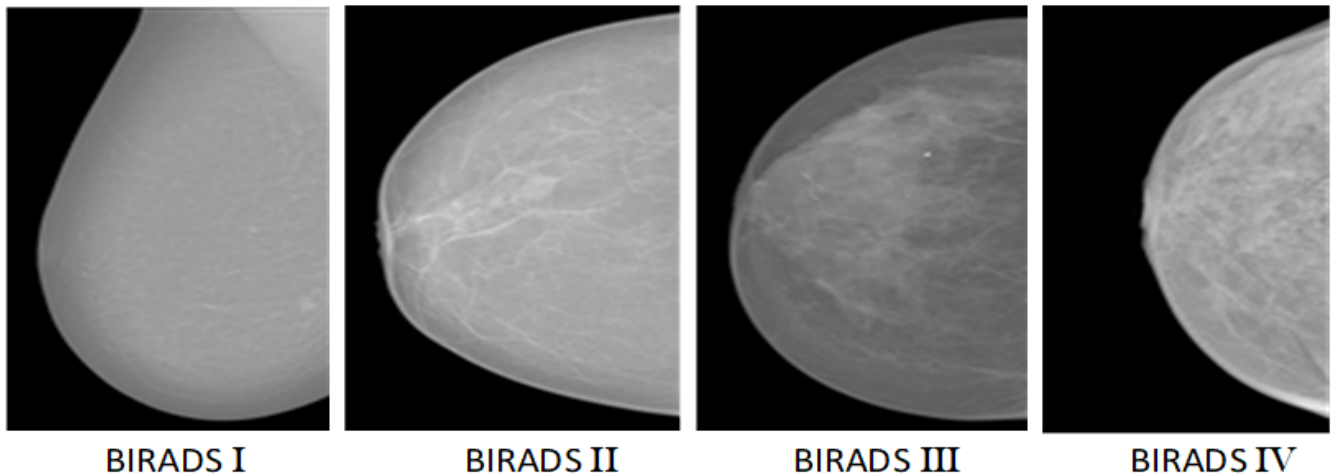**FIGURE 2.** Edges of breast mass [6].

FIGURE 3. Mammogram images example of four BI-RADS breast density category ('BI-RADS.I'): 'Fatty', ('BI-RADS.II'): 'Fibroglandular', ('BI-RADS.III'): 'Hetrogenously', ('BI-RADS.IV'): 'Extremely dense.'

The American College of Radiology created 'BI-RADS,' which stands for Breast Imaging Data and Reporting System [9]. Figure 3 illustrates the 'BI-RADS evaluation categories. 'BI-RADS' advantageously encourages radiologists to consider which category is most appropriate.

B. COMPUTER-AIDED DIAGNOSIS (CAD)
A computer-aided diagnostic (CAD) system analyzes radiographic evidence to determine the likelihood that a feature represents a certain disease process.'' (eg benign vs. malignant) [10]. CAD systems for breast cancer utilize various pattern recognition techniques.

In general, there are three main modules in a CAD system: mammogram, the shape of a specific breast mass can be 'Round,' 'Oval,' 'Lobular,' 'Irregular,' or 'architectural distortion' as shown in Figure 1.

'Circumscribed oval'- and 'round-shaped' masses strongly suggest that a lesion is benign. In contrast, masses of irregular shape usually raise suspicion of malignancy, the edges of breast detection, segmentation, and classification. Mass detection is a challenging issue but plays a significant role in the diagnosis of breast cancer. The detection task is to find the location of a lesion on a mammogram if one exists. Detection generally comprises three modules: (1) detecting suspicious regions (i.e., by density, micro-calcifications, and mass),

(2) extracting features, and (3) eliminating false positive regions. Mass segmentation is the next stage, which is the process of partitioning mammogram images into regions possessing identical characteristics. Mass classification is the last stage, categorizing the input regions of interest (ROIs) as Mass or Normal, depending on abnormality. Mass lesions are then categorized as benign or malignant. Breast mass classification can be grouped for the data training stage. Breast mass detection and breast density classification would greatly help treat breast cancer. One problem with using CAD systems for mass detection is the high false-positive rate since masses and normal dense tissue is comparable on a mammogram.

III. LITERATURE REVIEW

Many automatic systems for breast cancer classification have emerged in recent years; these systems use different approaches. Breast cancer categorization is a classification problem that requires the extraction of discriminatory features and then classification. State-of-the-art strategies for breast cancer staging that have been proposed are discussed in the following paragraphs.

Whitaker *et al.* [10] suggested a two-stage patch classification technique for mammography using two texture descriptors: "Histogram of Oriented Texture (HOT)" and "Pass Band Discrete Cosine Transform (PB-DCT)." In the first stage, mammogram patches are classified as normal or abnormal. The second stage uses a support vector machine (SVM) to classify aberrant mammographic regions as benign or malignant. Jothilakshmi and Raaza [13] developed a texture-based strategy to identify malignant and benign using multiple SVMs, with features retrieved using "grey-level co-occurrence matrices (GLCM). In [14] proposed a new approach to classify benign and malignant breast masses. The approach converts two-dimensional contours of breast masses on mammography into a one-dimensional signature.

DT is a popular classification method that is easy to learn and interpret while improving human readability. On the diagnostic data set, the researchers used 10-fold cross-validation to discover the optimal combination of parameters, and this model had an accuracy of 93.62% and a specificity of 90.66% [22]. In a multidimensional feature space filled with known instances of a training dataset, nearest neighbor algorithms classify the data by discovering its nearest neighbors [23]. The better the dimension ratios for the nearest neighbors, the higher the predictive efficiency. Because the results of this algorithm depend on how the distance between the data is measured, both approaches, the Manhattan distance, and the Euclidean distance were examined this time. SVM is a supervised learning technique for classifying, predicting, and detecting outliers. They are cheap and effective, particularly in high-dimensional areas, because they only require a subset of training points on support vectors. On diagnostic data, SVM achieved 98 percent accuracy [24] and 78.35 percent accuracy using the polynomial kernel.

Since NNs are capable of capturing the relationships between attributes, it is largely used for BC detection. Liu *et al.* [25] presented a DT algorithm for BC detection and used an under-sampling approach to address the issue of an imbalanced training class, which improved the results. Quinlan [26] presented a better DT method for BC detection and achieved a performance accuracy of 94%. However, one classifier cannot learn all the features of the BC detection and recurrence rate [27]. Considering the drawback of a single classifier, various ensemble-based algorithms are proposed. Akay [28] used the hybrid method proposed by Zhang *et al.* [29] where the authors presented a hybrid classifier with various sets of features and used SVM for classification. The fuzzy approach for feature selection and the fuzzy nearest neighbor method for BC detection are combined into a hybrid classification system for BC detection [24].

Data can be effectively classified using classification and data mining methods. Such methods have prevalent use in the medical field for analysis and diagnosis so decisions can be made. Classification techniques [30] such as AdaBoost, KNN, and K Tree, as well as neural networks, feature selection methods, and SVM [29], have been used in a variety of study domains. Goodman *et al.* [31] used three different approaches, including artificial immune recognition system (AIRS), optimized learning vector quantification (LVQ), and large LVQ, with an accuracy of 97.2%, 96.7%, and 96. 8%, respectively. Using the SVM class possibility based kernel (CPBK) algorithm, Li and Liu [32] achieved a classification accuracy of 93.26%. The accuracy of an SVM-based classifier reported in [33] was 97.60%.

Ensemble of classifiers is another essential method to improve the performance of a single classifier [34]. The predictions by a single classifier are combined by various techniques in the ensemble-based classifiers, which improves the overall prediction and makes more accurate predictions than a single classifier [35]. In typical settings of the ensemble-based classifiers, training data is replicated by k times and then build k classifiers by re-sampling the original data [36]. Similarly, various voting methods exist for classification. To make the qualifier run multiple times, take a training set and a qualifier and change the distribution of instances in the training set. To arrive at the final classification, the results of the constructed classifiers [37] are concatenated. The most prevalent voting approach is to combine the results of the base-level classifiers using a plurality. However, this strategy does not use metal fetching, and all training sets and classifiers use the same voting technique [38].

The one-dimensional signature is then segmented into subsections to extract local contour features. Finally, these features are fed to an SVM classifier. Laroussi *et al.* [39] Proposed two CAD systems for the classification of mammograms breast density for two and four "BI-RADS" classes consisting of features computed using different Law filters of varying lengths. The feature vectors are then fed to classifiers 'PNN,' 'NFC,' and 'SVM' to classify tissue density. The literature reports several studies indicating the utility of

TABLE 2. Summary of breast cancer classification.

Reference	Problem	Features	Classifier	Dataset	Accuracy
Shastri et al. [11]	mass vs. Normal	(HOT)	(SVM)	MIAS/150 DDSM/2576	MIAS overall acc 97.33% DDSM overall acc 84.84%
	Benign vs. Malignant	(PB-DCT)			MIAS overall acc100% DDSM overall acc 72.43%
Chokri and Farida [12]	Benign vs. Malignant	(SGLDM)	(MLP)	DDSM/480	88.02%
		(RDM)			
		Textural and intensity			
Jothilakshmi et al. [13]	Benign vs. Malignant	(GLCMs)	(SVM)	MIAS/50	94%
Li et al. [14]	Benign vs. Malignant	intensity histogram, shape, and texture	(ANN) (SVM) (KNN)	DDSM/323	99.66%
Gedik [15]	abnormal vs. Normal	shearlet transform, t-test statistic and dynamic thresholding	(SVM)	MIAS/228	98.29%
				DDSM/ 228	96.89%
	Benign vs. Malignant			MIAS/228	97.39%
				DDSM/ 228	98.08%
Chisako et al. [16]	Benign vs. Malignant	(RLTP)	(ANN) (SVM) (RF)	private/376 ROIs	not given
		(GLCMs)			
Lima et al. [17]	malignant vs. benign vs. normal	wavelets and Zernike moments	(SVM)	IRMA/355	94.11%
Rahimeh et al. [18]	Benign vs. Malignant	statistical and (GLCMs)	(MLP)	MIAS/57	90.94%
Rabidas et al. [19]	Benign vs. Malignant	statistical (GLCMs)	(LDA) (SVM)	MIAS/58 DDSM/200	87.28% 75.60%
Yanfeng Li et al. [20]	Benign vs. Malignant	Texton	(KNN)	DDSM/114	85.96%
Beura et al. [21]	abnormal vs. Normal	Multiresolution analysis using (2D-DWT) and (GLCMs)	(BPNN)	MIAS/322	98%
				DDSM/ 550	98.80%
	Benign vs. Malignant			MIAS/115	94.20%
				DDSM/ 250	97.40%

hand-crafted features for breast cancer classification. Previous works have shown promising results using various classifiers. However, previous studies have demonstrated that using an ensemble of classifiers improves the results. In this work, we address this gap and present a method that uses different classifiers, i.e., an ensemble of classifiers, to improve the results of breast cancer detection. An overview of techniques is given in Table 2 for breast cancer classification methods.

IV. METHODOLOGY

This section presents the method used for an ensemble of ML classifiers. This architecture is composed of four different ML models. They are stacked and then further trained as an ensemble. After training, the ANN model is used for the outcome. The illustration of our proposed DL network is shown in Figure 4. The performance is compared with the several ML classifiers individually with and without up-sampling techniques. We also compared the performance of the proposed ensemble model with other ensemble models.

A. PROPOSED APPROACH

In this study, we design a classification framework by an ensemble of four ML-based classifiers named SVM, LR, NB, and DT. An ensemble model is stacked, and predictions

are concatenated and then fed to the ANN model for final prediction. Each of the algorithms used in our study is also briefly explained next section. The steps of the proposed model can be summarized below:

1) We used machine learning based classifiers on a training dataset
2) In the second step, the K-fold method retrieves the most common outcome from these classifiers.
3) In third step, we concatenated results from machine learning classifiers
4) New training dataset streamlined as a result
5) In this step, we input the new dataset into the default ANN
6) Result and evaluation of outputs

1) SUPPORT VECTOR MACHINE (SVM)

A supervised ML-based technique, SVM selects the moderate number of samples called support vectors and builds a linear discriminant function. SVM solved the restriction of linear limits [40]. SVM can be considered a two-class data set that can be partitioned linearly to show a maximum hyperplane margin. The new samples are linearly fit or appear linearly separable in the high-level plane following the selection of the appropriate mapping. The SVM tries to find the
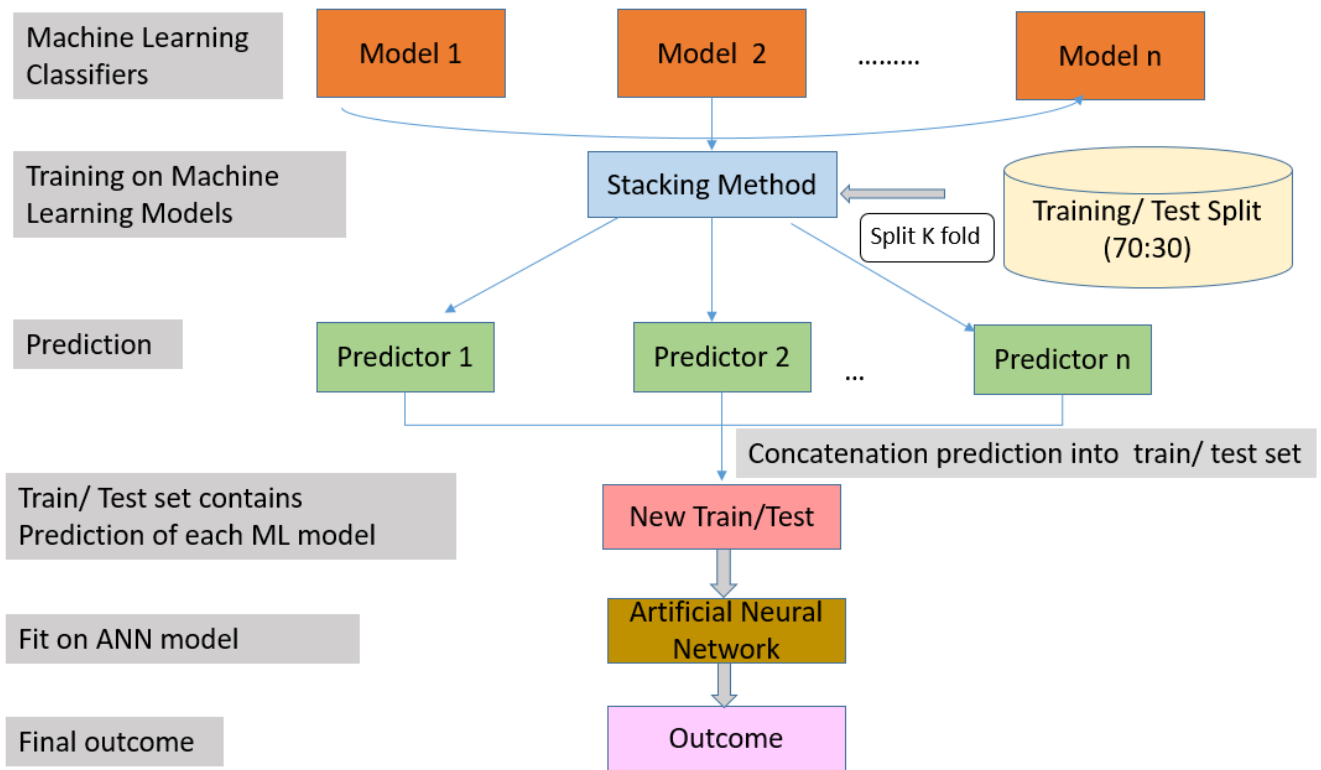
FIGURE 4. Illustration of the proposed network architecture.

most advantageous hyperplane that minimizes the distance between two groups [41].

2) LOGISTIC REGRESSION (LR)
The LR method is created by replicating the posterior probability of K groups across linear roles in x while ensuring they equal one and stay within the range [0, 1]. Logit shifts K -1, or log probabilities can be used to describe LR. Although the final group is used as the denominator in the odds ratio, the choice of the denominator is indeterminate because the counts are divided evenly. Since there is only one linear role when K = 2, the style is direct. This technique is often used in biostatic tasks where binary responses are repeated.

3) NAIVE BAYES (NB)
Bayes theorem [42] is used to suggest the NB algorithm. The NB classifier can be revised in the following ways using Bayes' theorem and the exact procedures [43]. We conclude that there is a training set of instances T. There are group marks on these specimens. $C_1, C_2 \cdots C_k$ are the names of the groups. Each specimen is an n-dimensional agent represented by the formula $X = x_1, x_2 \cdots x_n$. It states that X has n features since it has n dimensions. A specimen X is predicted to be a member of groupC_i if the probability that group i depends on X is greater than the probability that each of the other groups depends on X, or formally:

$$p(C_i|X) > P(C_j|X) \quad for \quad 1 \leq j \geq k_j = i) \quad (1)$$

Using Bayes' Theorem P (Ci |X) is calculated as follows:

$$P(C_i|X) = P(X|C_i)P(C_i)/P(X) \quad (2)$$

4) DECISION TREE (DT)
DT that begins with huge groupings of specimens within clearly defined categories [44]. Specimens are used for patterns that allow groups to be accurately characterized by combining nominal and numerical features. These markers are then represented as models, resulting in decision frameworks or sets of if-then processes that can be used to distinguish new samples, emphasizing making designs understandable and accurate. To determine the 'goodness' of a test, the C4.5 calculus uses equations based on theoretical data; specifically, they choose the test that takes the most data from a collection of specimens while limiting themselves to evaluating a single characteristic.

The limitation of DT is how to handle the issue of over-fitting and unknown values. C4.5 method of DT can address the issue of unknown values, especially the samples with unknown values are ignored. A classifier that categorizes all samples in the training data may not be as effective as a DT. To circumvent this, C4.5 uses an error rate-based pruning mechanism for all subtrees, and the subtree is removed when the computed error is raw. This strategy is more effective and produces better results [44].

5) ARTIFICIAL NEURAL NETWORK (ANN)

In the past decades, ANNs have been utilized by researchers, thus making them a relevant research area. Greatly, the network has enabled great success, especially in BC classification and early-stage prognosis [45]. ANN models usually have three layers: input, hidden, and output [31]. The layers comprise interconnected neurons with nonlinear switching activation functions to enhance nonlinear capacity. First, the input layer gets the data, then passes it to a hidden layer for analysis and returns the results to the output layer. Results shows are now displayed through the output layer. However, given the constraints, training an ANN will likely require long informal chains of computing processes. There are three dense layers and two dropout levels in the ANN structure used in this study. The DNN, on the other hand, is made up of five dense layers and three dropout layers.

V. EXPERIMENTAL RESULTS

In this section, we present the datasets used in this study and the experimental evaluations to demonstrate the usefulness of our proposed model. In this study, following previous studies, we used accuracy to evaluate the performance. Classification results are analyzed using a 10-fold cross-validation technique.

6) DATASET DETAILS

The Breast cancer Wisconsin (Diagnosis)[1] and Breast cancer Wisconsin (Prognosis)[2] databases are used in this study.

TABLE 3. Dataset distribution.

Diagnosis		Prognosis	
malignant	benign	non-recur	recur
2	357	151	47

Wisconsin Breast Cancer (Diagnosis) contains 569 instances and 32 attributes (an ID and a target variable). Wisconsin Breast Cancer (Prognosis) contains 198 instances and 34 attributes (containing an ID and a target variable). The forecast dataset also had four missing attribute values, which were removed; furthermore, the forecast data set is considerably skewed, with 151 non-recurring and 47 recurring outcomes. Dataset distribution is given in Table 3. In the BC Wisconsin Diagnostic and Prognostic data sets, two additional strategies (algorithm approach and data approach) were implemented to solve the problem of an unbalanced classification problem. To start with, we used cost-sensitive learning or a misclassification penalty as a misclassification penalty while training the model to improve performance in minority classes.

---

[1]https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[2]https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(Prognostic)

This is achieved by adding the misclassification cost to the error or using it to weight the error. Second, the dataset can be resampled, making this approach more versatile. Upsampling and downsampling are used to increase the number of minority classes in the resampling. Data standardization was done to ensure that the data was consistent. Each type of data had the same content and format.

7) EVALUATION METRIC

We have used accuracy as a evaluation metric is this study. All reported results are the average results. Accuracy is defined as follows:

$$Accuracy : \frac{Number of correct predictions}{All samples} \quad (3)$$

8) EXPERIMENTAL RESULTS AND DISCUSSION

This section describes the results and presents the baseline models used to compare our proposed model. To evaluate the performance of the proposed model, we compare it with various machine learning classifiers and other methods. All the algorithms were used with their default parameters in this study. Table 4 and Table 5 show the results of all ML classifiers and DL classifiers used in our study.

TABLE 4. Comparison of ML classifiers. Average accuracy (%) is reported as evaluation metric.

Machine Learning Models	Diagnosis	Prognosis
Decision Tree	91.22	76.26
Random Forest	97.07	76.26
Logistic Regression	98.00	75.24
Support Vector Machine	98.10	78.35
Naive Bayes	93.56	70.71

TABLE 5. Comparison of DL classifiers.

Deep Learning Models	Diagnosis	Prognosis
Deep Neural Network	95.36	87.5
Artificial Neural Network	98.24	90.22

It can be observed from Table 4 that SVM outperforms all ML-based classifiers on both diagnosis BC with 98.10% accuracy and prognosis BC dataset with 78.35% accuracy. In contrast, the worst classifier on the diagnosis dataset is DT with 91.22% accuracy and NB on the prognosis dataset with 70.71%. For DL-based classifiers, ANN performs well on both datasets, with 98.24% accuracy on the diagnosis dataset and 90.22% accuracy on the prognosis dataset compared to DNN, which can be seen in Table 5. We also tested a combination of ML and DL-based models. Furthermore, to address the problem caused by uneven data distribution and small sample size, each model combination is upsampled and the results compared, as shown in Table 6.

TABLE 6. Ensemble of ML classifiers with ANN with and without samplying.

Model	Diagnosis		Prognosis	
	Without Sampling	With Sampling	Without Sampling	With Sampling
(SVM+LR+NB+DT)+ANN	97.67%	98.83%	81.35%	84.70%
(SVM+LR+NB+RF)+ANN	97.07%	98.24%	83.05%	88.13%
(SVM+LR+RF+DT)+ANN	97.66%	98.24%	83.00%	84.74%
(SVM+LR+RF+NB) + ANN	97.07%	98.24%	83.15%	88.33%
(SVM+LR+RF) + ANN	95.91%	98.14%	81.36%	77.96%
(SVM+LR) + ANN	96.46%	96.46%	76.27%	76.27%

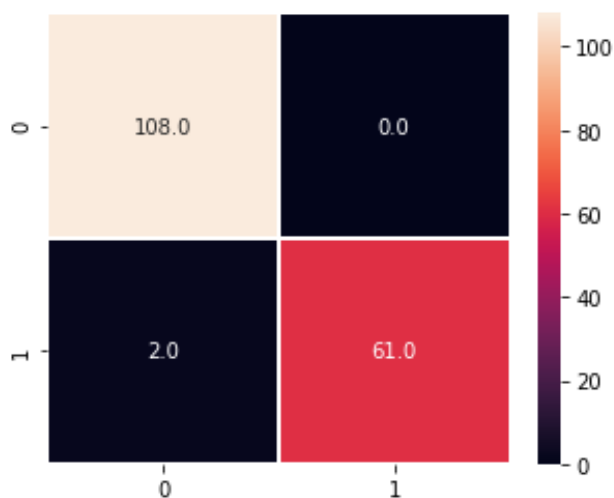**FIGURE 5.** Illustration of confusion matrix (Diagnosis).

FIGURE 6. Illustration of train/test accuracy (Diagnosis).

Comparing all the different ensemble models in Table 6 shows which model performed best for each dataset. The best ensemble model is the ensemble of (SVM + LR + NB + DT) in both cases (without 97.67% and with upsampling 98.83%). In contrast, the worst-performing combination is (SVM+LR+RF) in both cases (95.91% for without sampling and 98.14% for upsampling) on the diagnosis dataset. For prognosis, the best ensemble model combines (SVM+LR+RF+NB) in both cases (without 83.15% and with upsampling 88.33%). In contrast, the worst-performing combination is (SVM+LR) in both cases (76.27% for without and 76.27% with up-sampling) on prognosis. The increment of 1.16% was observed on diagnosis and 5.18% on the prognosis dataset when the upsampling technique was used. The confusion matrix and train/test accuracy of best-performing ensemble classifiers can be seen in Figure 5 and Figure 6, respectively.

9) ANALYSIS
We also analyzed the effects of applying balanced class weights with sampling and measured the performance when applying different K values, as shown
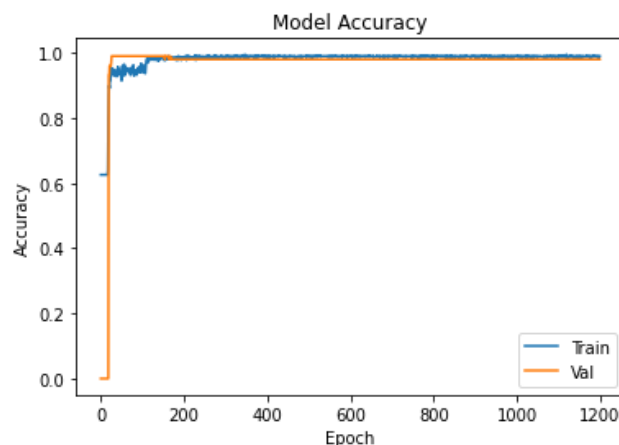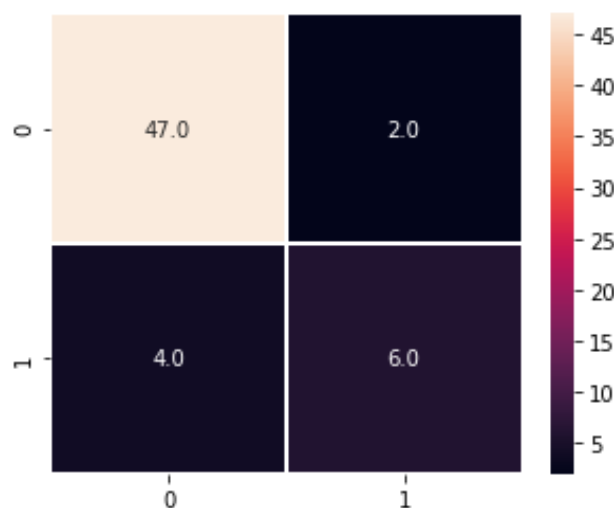

FIGURE 7. Illustration of confusion matrix (Prognosis) when K = 5.

in Figure 7 and Figure 8. We observed that performance increased substantially for all tested combinations of classifiers when compared with upsampling on the prognosis dataset. We also note that the confusion matrices below show
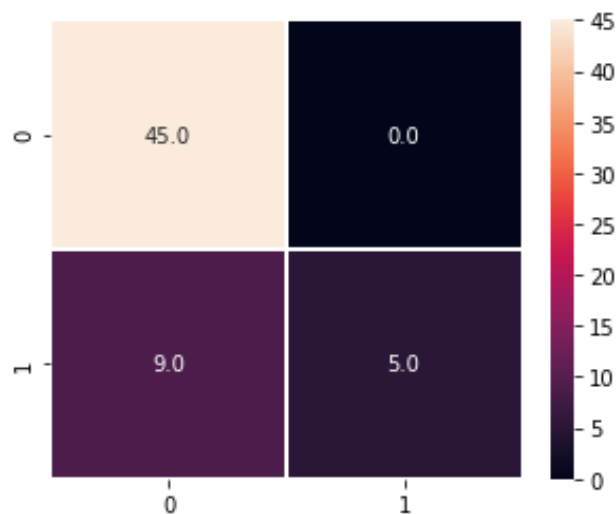
FIGURE 8. Illustration of confusion matrix (Prognosis) when K = 10.

that when K is 5 instead of 10, the model (SVM + LR + RF + DT) trained on the forecast outperforms.

VI. CONCLUSION

We proposed a method for breast cancer diagnosis and prognosis using machine learning techniques in this research. Benchmark datasets are used for the experiments. Classifiers based on machine learning and deep learning have shown their exceptional potential to increase classification and prediction accuracy. Several ensembles of different ML-based classifiers were also tested for the classification of BC. We found out that SVM outperforms both datasets compared to all ML classifiers and ANN from DL classifiers when used individually. For the ensembling method, (SVM + LR + NB + DT) performs well without and with upsampling on the diagnosis dataset, whereas (SVM+LR+RF+NB) outperforms all other combinations on the prognosis dataset when ANN is used as a final layer. We also observed an increase in performance when balanced class weights are used along with the upsampling technique as compared to without, and the upsampling technique is used individually. The performance was also analyzed using a different number of K-fold for the best ensemble classifier. In the future, we intend to apply more advanced models for the automatic detection of BC.

REFERENCES

[1] *Cancer Burden Rise to 18.1 Million New Cases and 9.6 Million Cancer Deaths in 2018*, International Agency for Research on Cancer, World Health Organization, Geneva, Switzerland, 2018, p. 3.
[2] W. S. Alqahtani, N. A. Almufareh, D. M. Domiaty, G. Albasher, M. A. Alduwish, H. Alkhalaf, B. Almuzzaini, S. S. Al-Marshidy, R. Alfraihi, A. M. Elasbali, and H. G. Ahmed, "Epidemiology of cancer in Saudi Arabia Thru 2010–2019: A systematic review with constrained meta-analysis," *AIMS Public Health*, vol. 7, no. 3, p. 679, 2020.
[3] J. Wei, H.-P. Chan, Y.-T. Wu, C. Zhou, M. A. Helvie, A. Tsodikov, L. M. Hadjiiski, and B. Sahiner, "Association of computerized mammographic parenchymal pattern measure with breast cancer risk: A pilot case-control study," *Radiology*, vol. 260, no. 1, pp. 42–49, Jul. 2011.
[4] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. V. D. Laak, B. V. Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[5] M. Milosevic, D. Jankovic, A. Milenkovic, and D. Stojanov, "Early diagnosis and detection of breast cancer," *Technol. Health Care*, vol. 26, no. 4, pp. 729–759, 2018.
[6] B. Surendiran and A. Vadivel, "Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer," *Int. J. Med. Eng. Informat.*, vol. 4, no. 1, pp. 36–54, 2012.
[7] R. A. Castellino, "Computer aided detection (CAD): An overview," *Cancer Imag.*, vol. 5, no. 1, pp. 17–19, 2005.
[8] E. Darcey, N. McCarthy, E. K. Moses, C. Saunders, G. Cadby, and J. Stone, "Is mammographic breast density an endophenotype for breast cancer?" *Cancers*, vol. 13, no. 15, p. 3916, Aug. 2021.
[9] W. A. Berg, C. Campassi, P. Langenberg, and M. J. Sexton, "Breast imaging reporting and data system: Inter- and intraobserver variability in feature analysis and final assessment," *Amer. J. Roentgenol.*, vol. 174, no. 6, pp. 1769–1777, Jun. 2000.
[10] K. Whitaker, "Earlier diagnosis: The importance of cancer symptoms," *Lancet Oncol.*, vol. 21, no. 1, pp. 6–8, Jan. 2020.
[11] A. A. Shastri, D. Tamrakar, and K. Ahuja, "Density-wise two stage mammogram classification using texture exploiting descriptors," *Expert Syst. Appl.*, vol. 99, pp. 71–82, Jun. 2018.
[12] F. Chokri and M. H. Farida, "Mammographic mass classification according to Bi-RADS lexicon," *IET Comput. Vis.*, vol. 11, no. 3, pp. 189–198, Apr. 2017.
[13] G. R. Jothilakshmi and A. Raaza, "Effective detection of mass abnormalities and its classification using multi-SVM classifier with digital mammogram images," in *Proc. Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, Jan. 2017, pp. 1–6.
[14] H. Li, X. Meng, T. Wang, Y. Tang, and Y. Yin, "Breast masses in mammography classification with local contour features," *Biomed. Eng. OnLine*, vol. 16, no. 1, pp. 1–12, Dec. 2017.
[15] N. Gedik, "A new feature extraction method based on multi-resolution representations of mammograms," *Appl. Soft Comput.*, vol. 44, pp. 128–133, Jul. 2016.
[16] C. Muramatsu, T. Hara, T. Endo, and H. Fujita, "Breast mass classification on mammograms using radial local ternary patterns," *Comput. Biol. Med.*, vol. 72, pp. 43–53, May 2016.
[17] S. M. L. de Lima, A. G. da Silva-Filho, and W. P. dos Santos, "Detection and classification of masses in mammographic images in a multi-kernel approach," *Comput. Methods Programs Biomed.*, vol. 134, pp. 11–29, Oct. 2016.
[18] R. Rouhi and M. Jafari, "Classification of benign and malignant breast tumors based on hybrid level set segmentation," *Expert Syst. Appl.*, vol. 46, pp. 45–59, Mar. 2016.
[19] R. Rabidas, J. Chakraborty, and A. Midya, "Analysis of 2D singularities for mammographic mass classification," *IET Comput. Vis.*, vol. 11, no. 1, pp. 22–32, Feb. 2017.
[20] Y. Li, H. Chen, G. K. Rohde, C. Yao, and L. Cheng, "Texton analysis for mass classification in mammograms," *Pattern Recognit. Lett.*, vol. 52, pp. 87–93, Jan. 2015.
[21] S. Beura, B. Majhi, and R. Dash, "Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer," *Neurocomputing*, vol. 154, pp. 1–14, Apr. 2015.
[22] R. AlamKhan, N. Ahmad, and N. Minallah, "Classification and regression analysis of the prognostic breast cancer using generation optimizing algorithms," *Int. J. Comput. Appl.*, vol. 68, no. 25, pp. 42–47, Apr. 2013.
[23] R. Nisbet, J. Elder, and G. D. Miner, *Handbook of Statistical Analysis and Data Mining Applications*. New York, NY, USA: Academic, 2009.
[24] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 6844–6852, 2015.
[25] Y.-Q. Liu, C. Wang, and L. Zhang, "Decision tree based predictive models for breast cancer survivability on imbalanced data," in *Proc. 3rd Int. Conf. Bioinf. Biomed. Eng.*, Jun. 2009, pp. 1–4.
[26] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *J. Artif. Intell. Res.*, vol. 4, pp. 77–90, Mar. 1996.
[27] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, Jun. 2018.

[28] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3240–3247, 2009.

[29] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, pp. 1476–1482, Mar. 2014.

[30] A. G. Dunn, J. Leask, X. Zhou, K. D. Mandl, and E. Coiera, "Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: An observational study," *J. Med. Internet Res.*, vol. 17, no. 6, p. e144, Jun. 2015.

[31] D. E. Goodman, L. Boggess, and A. Watkins, "Artificial immune system classification of multiple-class problems," in *Proceedings of the Artificial Neural Networks in Engineering ANNIE*, vol. 2. St. Louis, MO, USA: ASME, 2002, pp. 179–183.

[32] D.-C. Li and C.-W. Liu, "A class possibility based kernel to increase classification accuracy for small data sets using support vector machines," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3104–3110, Apr. 2010.

[33] H. Cevikalp, B. Triggs, H. S. Yavuz, Y. Küçük, M. Küçük, and A. Barkana, "Large margin classifiers based on affine hulls," *Neurocomputing*, vol. 73, nos. 16–18, pp. 3160–3168, Oct. 2010.

[34] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, nos. 1–2, pp. 1–39, Feb. 2010.

[35] T. G. Dietterich, "Machine-learning research," *AI Mag.*, vol. 18, no. 4, p. 97, 1997.

[36] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[37] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Mach. Learn.*, vol. 36, nos. 1–2, pp. 105–139, 1999.

[38] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?" *Mach. Learn.*, vol. 54, no. 3, pp. 255–273, Mar. 2004.

[39] M. G. Laroussi, N. G. B. Ayed, A. D. Masmoudi, and D. S. Masmoudi, "Diagnosis of masses in mammographic images based on Zernike moments and local binary attributes," in *Proc. World Congr. Comput. Inf. Technol. (WCCIT)*, Jun. 2013, pp. 1–6.

[40] G. Meng and H. Saddeh, "Applications of machine learning and soft computing techniques in real world," *Int. J. Comput. Appl. Inf. Technol.*, vol. 12, no. 1, pp. 298–302, 2020.

[41] V. Vapnik, *The Nature of Statistical Learning Theory*. Cham, Switzerland: Springer, 1999.

[42] R. R. Yager, "An extension of the naive Bayesian classifier," *Inf. Sci.*, vol. 176, no. 5, pp. 577–588, Mar. 2006.

[43] W. Zheng, X. Zhu, Y. Zhu, R. Hu, and C. Lei, "Dynamic graph learning for spectral feature selection," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29739–29755, 2018.

[44] J. R. Quinlan, *The Morgan Kaufmann Series in Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[45] D. Furundzic, M. Djordjevic, and A. J. Bekic, "Neural networks approach to early breast cancer detection," *J. Syst. Archit.*, vol. 44, no. 8, pp. 617–633, Apr. 1998.

JUNAID RASHID received the B.S. and M.S. degrees in computer science from the COMSATS Institute of Information Technology, Wah Campus, Pakistan, in 2014 and 2016, respectively, and the Ph.D. degree from the Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan, in 2020. He was a Visiting Lecturer at COMSATS University Islamabad, Wah, from September 2015 to January 2016, from September 2016 to January 2017, and from September 2019 to February 2020. From March 2017 to January 2020, he was a Research Associate and a Lecturer at the University of Engineering and Technology. From January 2020 to July 2021, he worked as a Lecturer with the Department of Computer Science, Air University Islamabad, Kamra Campus, Pakistan. He has been working as an Honorary Senior Postdoctoral Research Fellow at the Center for AI and Data Science, Edinburgh Napier University, U.K., since November 2020, and awarded the Fellowship, in 2022. He is currently a Postdoctoral Researcher at Kongju National University, South Korea. He has published many research papers in prestigious journals and conferences. His research interests include data science, machine learning, natural language processing, topic modeling, text mining, information retrieval, big data, pattern recognition, fuzzy systems, medical informatics, biomedical text analytics, and software engineering. He received a fully-funded scholarship for his Ph.D. degree. He has been working as an Academic Editor in *PLOS ONE*. He has served/been serving as a reviewer for various reputed journals and conferences.

LIAQAT ALI received the Ph.D. degree in information and communication engineering from the University of Electronic Science and Technology of China (UESTC), China. He is currently working as an Assistant Professor with the Department of Electrical Engineering, University of Science and Technology Bannu, Pakistan. His research interests include biomedical and health informatics, machine learning, data science, and soft computing. More specifically, his research is concentrated on the development of automated disease detection (decision making/prediction) systems by using novel feature selection and extraction methods for improving the performance of conventional machine learning models. Most of his research work has been published in top ranked (mostly Q1 ranked) journals. Additionally, he is an Active Reviewer of many top-tier journals, including *Expert Systems With Applications* (Elsevier), *Artificial Intelligence in Medicine*, *IRBM* journal, and *Medical and Biological Engineering and Computing*. He is serving as a Lead Editor for a Special Issue Open for Submission in *Journal of Healthcare Engineering*.

USMAN NASEEM received the B.Sc. degree in electrical engineering and the master's degree in analytics (research). He is currently a Lecturer and a Unit Coordinator at Sydney International. Before working in academia, he worked in leading ICT companies for more than ten years in various roles. He has published many peer-reviewed academic papers in well-reputed conferences and journals in this area as an active researcher. His research interests include natural language processing and data analytics in general, particularly in social media and healthcare analysis. As an academician, he is an Active Member of different professional bodies, such as the Association for Computing Machinery (ACM). He has served as the session chair/co-chair/TPC member for dozens of conferences. He is also an academic reviewer of several reputed journals.

JUNGEUN KIM received the Ph.D. degree in knowledge service engineering from KAIST. He is currently an Assistant Professor with Kongju National University (KNU). Before joining KNU, he was a Senior Researcher with the Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute (ETRI). His research interests include data mining, artificial intelligence, big data analysis with distributed processing platforms, and open data platforms.

QAZI EMAD UL HAQ received the Ph.D. degree from King Saud University, Riyadh, Saudi Arabia. He is currently working as an Assistant Professor at Naif Arab University for Security Sciences (NAUSS), Riyadh. His research interests include security analytics, cyber security, network security, deep learning, artificial intelligence, machine learning, and pattern recognition.

MUHAMMAD IMRAN is currently working as a Senior Lecturer with the School of Engineering, Information Technology (IT), and Physical Sciences, Federation University, Australia. He has published more than 300 research papers in reputable international conferences and journals. His research interests include mobile and wireless networks, the Internet of Things, cloud and edge computing, and information security. His research is supported by several grants. He serves as an Associate Editor for many top ranked international journals, including *IEEE Network* magazine, *Future Generation Computer Systems*, and IEEE Access.

• • •

MAZHAR JAVED AWAN received the M.S. degree in computer science from the University of Central Punjab (UCP), Lahore. He is currently pursuing the Ph.D. degree with Universiti Teknologi Malyasia (UTM). He is an Assistant Professor with the Department of Software Engineering, University of Management and Technology (UMT), Lahore. He has overall 17 years of teaching experience in various institutes. He has published in high standard ISI journals, such as IEEE Access, *Diagnostics*, and *Computers, Materials and Continua* (CMC). His research interests include data sciences, big data analytics, deep learning in medical images, natural language processing, and machine learning. Besides research, he is also a keynote speaker at national and international level at various conferences and workshops related to data science and big data. He is an IEEE Professional Member and an Advisor of UMT Branch of IEEE ComSoc. He has reviewed many WOS and Scopus Indexed journals, like IEEE INTERNET OF THINGS and *CMC* journals.