IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

Prediction of breast cancer comparative review of machine learning

Sravanthi Sallaram[1]
Research Scholar[1]
Computer Science and Engineering[1]
Koneru Lakshmaiah Education Foundation,
Vaddeswaram,AP,India[1]
mnrphd@gmail.com

Dr. M. Nageswara Rao[2]
Professor[2]
Computer Science and Engineering[2]
Koneru Lakshmaiah Education Foundation,
Vaddeswaram,AP,India[2]
Sravanthi.raaj@gmail.com

ABSTRACT

A tumour that grows in the tissues of the breast is called breast cancer. One of the leading causes of death for women worldwide, it is the most common type of cancer found in women. The methods of data mining, deep learning, and machine learning for breast cancer prediction are contrasted in this article. Breast cancer diagnosis and prognosis has been the subject of numerous studies; the accuracy rate of each approach varies according to the situation, tools, and datasets used. In order to identify the most appropriate method for handling massive datasets with high prediction accuracy, our main objective is to compare various current machine learning and data mining approaches.The primary objective of this review is to highlight all of the earlier research on machine learning algorithms used to predict breast cancer. This article gives novices who wish to analyse machine learning algorithms a thorough understanding of deep learning all the information they need.

1. INTRODUCTION

Among the most deadly and varied illnesses of our day, breast cancer claims the lives of many people all over the world. It is the second most common cause of death among women [1]. To forecast breast cancer, a variety of data mining and machine learning techniques are used [2]. One crucial difficulty is determining the most palatable and efficient algorithm for predicting breast cancer. Uncontrollably developing malignant tumours are the cause of breast cancer [3]. Breast cancer develops when several of the fatty and fibrous tissues of the breast start to grow abnormally.Different stages of cancer are brought on by cancer cells spreading throughout tumours. Breast cancer can be divided into various types and develops when damaged cells and tissues spread throughout the body [4].DCIS, sometimes referred to as non-invasive carcinoma, is a kind of breast cancer that arises when aberrant cells spread outside the breast [5]. Invading ductal carcinoma (IDC) [6], also known as infiltrative ductal carcinoma [7], is the second type. This type of cancer is more frequently found in men and arises when aberrant breast cells spread throughout the breast tissues [8]. Mixed tumour breast cancer (MTBC), sometimes referred to as invasive mammary breast cancer, is the third type of breast cancer [9]. Abnormal duct cells and lobular cells are the source of this type of cancer [10]. Lobular breast cancer (LBC), which arises inside the lobule, is the fourth type of cancer [11]. It increases the risk of developing more aggressive cancers.

Mucinous breast cancer (MBC) [12] is the fifth type of breast cancer generated by invasive ductal cells and is also referred to as colloid cancer. It happens whenabnormal tissue spreads across the duct [13]. The last type is inflammatory breast cancer (IBC), which causes swelling and redness of the breasts. It is a fast-growing breast cancer that starts when lymph tubes become clogged with damaged cells [14].Data mining is the process of obtaining meaningful information from big datasets.

Machine learning, statistics, databases, fuzzy sets, data warehouses, and neural networks are examples of data mining techniques and functions that help with the diagnosis and prognosis of many cancer disorders [15], including prostate cancer, lung cancer [16], and leukaemia [17].Traditional cancer screening approaches include "the gold standard" procedure, which comprises three tests: clinical evaluation, radiological imaging, and pathology [18]. This conventional method detects the presence of cancer via regression, whereas newer machine learning techniques and algorithms focus on model construction.

The model is designed to anticipate previously unknown data and achieves the projected outcomes during the training and testing stages [19]. The machine learning method is divided into three primary steps: preprocessing, feature selection or extraction, and classification [20].Feature extraction is the most crucial step in the machine learning process, and it helps in cancer diagnosis and prognosis. This approach is capable of distinguishing between benign and malignant tumours.

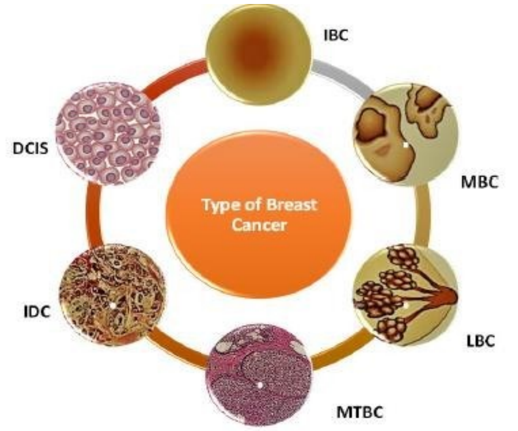
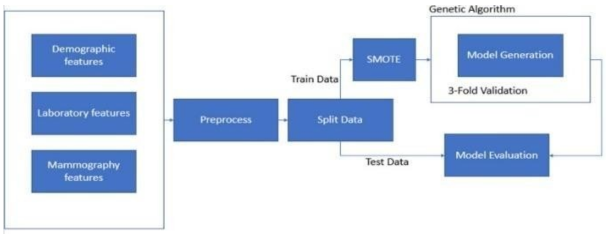


FIG 1. Display of the various types of breast cancer. Figure 1 shows how data mining and machine learning algorithms help in the detection and prediction of some types of breast cancer. Data mining techniques [22] such as classification, regression, and clustering help us get relevant information about breast cancer patients. These algorithms [23] have a training dataset that can be used to predict different forms of breast cancer [24].



Block diagram of methods

2. RELATED WORK

Breast cancer risk factors and prevention. Y.S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, et al., 2017 [1]. Breast cancer is the second biggest cause of cancer-related deaths among women. Breast cancer develops in a multi-step process involving various cell types, and prevention remains a challenge worldwide. Breast cancer prevention is best achieved by early detection. Early prevention has increased the 5-year relative survival rate of breast cancer patients in several developed nations to more than 80%. Over the last decade, significant progress has been achieved in the understanding of breast cancer as well as the development of prevention measures. Discovering breast cancer stem cells reveals the pathophysiology and tumour drug-resistance mechanisms, as well as several genes associated with breast cancer. People now have additional pharmacological alternatives for the chemoprevention of breast cancer, and biological prevention has recently been created to improvepatients' quality of life. In this review, we will summarise significant studies on breast cancer pathogenesis, associated genes, risk factors, and preventative measures conducted in recent years. These discoveries are a minor step in the protracted battle against breast cancer.

Y. Khourdifi and M. Bahaj (2018) used the finest machine learning algorithms to predict and classify breast cancer. The number of deaths caused by breast cancer is increasing dramatically every year. It is the most common type of cancer and the leading cause of mortality in women globally. Any advancement in the prediction and diagnosis of cancer is critical to living a healthy life. As a result, excellent cancer prediction accuracy is critical for updating patient therapy and survivability standards.

Machine learning techniques can make a significant contribution to the process of breast cancer prediction and early diagnosis; they have become a research hotspot and have been proven to be effective. In this study, we applied five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5), and K- Nearest Neighbours (KNN) on the Breast Cancer Wisconsin Diagnostic dataset. After obtaining the results, a performance evaluation and comparison is carried out between these different classifiers.

. The primary goal of this study work is to predict and diagnose breast cancer using machine learning algorithms and determine which are the most effective in terms of confusion matrix, accuracy, and precision. Support vector Machine surpassed all other classifiers, with the highest accuracy (97.2%).All work is carried out in the Anaconda environment, which is based on the Python programming language and the Scikit-learn library.

A review of the identification of breast cancer in medical imaging. Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu (2018)[3]. Breast cancer is among the most frequent types of cancer. Breast pathological image processing has become a major tool for early cancer detection. Using medical image processing to help doctors discover suspected breast cancer as early as feasible has long been a hot topic in medical image diagnostics. This study systematically describes a breast cancer recognition approach based on image processing in four steps: breast cancer detection, picture segmentation, image registration, and image fusion. The accomplishments and applications of supervised learning, unsupervised learning, deep learning, CNN, and other techniques in breast cancer detection are discussed. The possibility of using unsupervised and transfer learning to diagnose breast cancer is discussed.

Finally, breast cancer patients' privacy rights are discussed.Breast cancer is among the most common types of malignant tumours. Breast cancer is the most prevalent malignant tumour in Chinese women, according to a survey, and the incidence rate is increasing year after year. Early detection and therapy are critical for lowering breast cancer mortality.

Currently, mammography is the most routinely used tool for detecting breast cancer. However, due to the large volume of data and the poor imaging characteristics of early breast cancer, early detection is extremely challenging. With the advancement of image processing and early diagnosis technologies, image processing of breast pathology has become an essential method of early detection of breast cancer, which primarily involves the analysis of masses, calcifications, and breast density. Mass is one of the most common symptoms of breast cancer in mammography images. The main procedures in pathological image processing are as follows: first, image pre-processing, which involves eliminating background, markers, pectoral muscle, and noise, then breast segmentation and image enhancement.

Second, the region of interest is identified using a basic image processing method. Then, quality-representative features such as texture and morphological traits are retrieved. Finally, the tumour and normal tissue were separated using the extracted features. A high breast density is another indicator of breast cancer on X-ray pictures.

Forced labelling and domain adaptation are used to predict advanced ductal cancer in situ. Hou, M. A. Mazurowski, L. J. Grimm, J. R. Marks, L. M. King, C.

C. Maley, et al., 2020 [5]. The purpose of this research is to use supplementary classes to improve a predictive model whose performance is hampered by the frequent issues of a small number of primary cases, high feature dimensionality, and poor class separability. Our clinical objective is to use mammographic features to predict whether ductal carcinoma in situ (DCIS) discovered during a needle core biopsy will later be upstaged or revealed to include invasive breast cancer. Methods:

To improve the prediction of pure DCIS (negative) versus upstaged DCIS (positive) cases, this study examines the adjunctive roles of two related classes: atypical ductal hyperplasia (ADH), a non-cancerous type of breast abnormality, and invasive ductal carcinoma (IDC), with 113 computer vision-based mammographic features extracted from each case. To improve the baseline Model A's categorisation of pure vs. upstaged DCIS, we created three other strategies (Models B, C, and D) that use different methods of embedding features or inputs.Based on ROC analysis, the baseline Model A had an AUC of 0.614 (95% CI, 0.496-0.733). All three new models outperformed the baseline, with domain adaptation (Model D) coming out on top with an AUC of 0.697 (95% confidence interval, 0.595-0.797). In conclusion, we improved DCIS upstaging prediction performance by integrating two related pathology classes in different training phases. Significance: All three new embedding strategies for related class data beat the baseline model, revealing not only feature similarities between these different classes, but also the potential for enhancing categorisation by employing other related classes.

Efficacy of the multidisciplinary tumour board conference in gynaecologic oncology: A prospective research. B. Lee, K. Kim, J. Y. Choi, D. H. Suh, J. H. No, and H.-Y. Lee, 2017 [10]. Evidence suggests that multidisciplinary tumour board conferences (MTBCs) improve patient care for a variety of cancer types. However, few retrospective studies have looked into MTBC efficacy in individuals with gynaecologic malignancies. We conducted a prospective study to determine how MTBCs influence patient management in gynaecologic oncology. This prospective analysis comprised 85 consecutive cases of gynaecologic oncology MTBCs presented at our tertiary university hospital between January 2015 and April 2016.

The primary objective was the rate of treatment plan change, which covered both large and minor alterations. Major alterations were defined as the exchange, addition, or removal of a treatment method. All other modifications were minor, such as changes in intramodality or treatment time. The secondary objectives were change rates in diagnosis, diagnostic work-up, and radiological and pathological findings.

M. K. Gupta and P. Chandra, 2020 [15], provide a comprehensive study of data mining. Data mining is crucial in many human activities because it reveals previously undiscovered beneficial patterns (or knowledge). Data mining has become a vital task in a wide range of application sectors, including banking, retail, medical, and insurance, as well as bioinformatics. This report presents a detailed survey of data mining research trends in order to provide a comprehensive picture. This study offers a systematic and comprehensive overview of various data mining tasks and approaches. This study also discusses several real- world uses of data mining. This paper also discusses the obstacles and issues surrounding data mining research.

Leveraging machine learning to identify breast cancer,

A. Reddy, B. Soni, and S. Reddy 2020 [18]. India has seen 30% of all breast cancer cases in recent years, and this figure is expected to rise further. Breast cancer is diagnosed in India every two minutes, with one woman dying every nine minutes. Cancer patients can live longer lives if they are detected and diagnosed early on.

This research provides a unique method for detecting breast cancer that uses Machine Learning algorithms. To assess performance, the authors ran an experimental analysis on a dataset. In comparison to existing methods, the proposed method generated highly precise and efficient outcomes. Breast cancer (BC) is a malignant tumour that activates in breast cells. Tumours have the ability to spread throughout the body. BC is a worldwide disease that wreaks havoc on the lives of women aged 25 to 50. With the possible increase in the number of BC cases in India, the situation is concerning.

During the last five years, the survival rate for BC patients in the United States has been around 90%, while in India it has been around 60%. BC projections for India in 2020 estimate that the number could reach two million. Specialist doctors have identified hormonal, lifestyle, and environmental factors that may raise a person's risk of getting BC. Over 5%-6% of BC patients have been connected to gene mutations that occurred over the family's history. Other variables that contribute to BC include obesity, advancing age, and postmenopausal hormone abnormalities.

As a result, there is no way to avoid breast cancer, but early identification can dramatically improve outcomes. Furthermore, this can significantly cut treatment expenses. However, cancer signs might appear unexpectedly, making early detection challenging. Mammograms and self-breast testing must be used to detect any early abnormalities before the tumour progresses. The primary goal of this work is to present a novice method for detecting BC. This work conducts a thorough investigation of existing cancer detection methods and delivers very accurate and efficient results.

Z. Salod and Y. Singh, 2019 [19], "A protocol for comparing the performance of machine learning algorithms in breast cancer screening and detection." Background: Breast cancer is a well-known global crisis. In 2018, the World Health Organisation reported

2.09 million BC occurrences and 627,000 deaths worldwide. Mammography is the traditional approach for screening breast cancer in industrialised nations, although breast self-examination and clinical breast inspection are used in poor countries.

The gold standard for detecting breast cancer is a threefold assessment: i) clinical examination, ii) mammography and/or ultrasonography, and iii) Fine Needle Aspirate Cytology. However, the emergence of less expensive, effective, and noninvasive technologies of cancer screening and detection would be advantageous.Design and techniques: We suggest using eight machine learning algorithms: i) Logistic Regression; ii) Support Vector Machine; iii) K-Nearest Neighbours; iv) Decision Tree; v) Random Forest; vi) Adaptive Boosting; vii) Gradient Boosting; viii) eXtreme Gradient Boosting; and blood test results using BC Coimbra Dataset (BCCD) from the University of California Irvine online database to create models for BC prediction.

To ensure model robustness, we will use: i) stratified k- fold cross validation; ii) correlation-based feature selection (CFS); and iii) parameter adjustment. The models will be verified against BCCD validation and test sets for both full and reduced features. Feature reduction affects algorithm performance. Seven measures will be used to evaluate the model, including accuracy.The expected impact of the study on public health: The CFS, in conjunction with the best-performing model(s), can help identify crucial particular blood tests that point to BC, potentially serving as an essential BC biomarker. The best- performing model(s) may potentially be used to develop an Artificial Intelligence tool to help doctors with breast cancer screening and detection.

Breast cancer detection and prediction using machine learning and data mining techniques: A review (S. Eltalhi & H. Kutrani, 2019) [20]. Breast cancer is the second leading cause of death for women. Early detection of breast cancer will improve patients' survival rates. Data mining and machine learning have been extensively employed in the diagnosis and early detection of breast cancer. The purpose of this study is to assess the role of machine learning and data mining approaches in breast cancer detection and diagnosis.

The majority of these research focused on diagnosing and prognosing breast cancer with the WEKA instrument. Several research compared classification methods for breast cancer prediction, including Decision Tree, Naïve Bayes, and Artificial Neural Networks. Data mining and machine learning have been extensively employed in the diagnosis and prognosis of breast cancer. Furthermore, data mining and machine learning assist medical researchers in identifying correlations between variables and predicting disease outcomes using historical data.

Machine learning can be used to improve breast cancer detection and diagnosis, as well as to avoid overtreatment. It could also help with accurate decision- making. As a result, the purpose of this research is to examine the role of machine learning and data mining approaches in breast cancer detection and diagnosis.

Research on machine learning methods and feature extraction for time series. L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou, and D. Chen, 2017, [23]. The purpose of this study is to investigate the relationship between various machine learning methods and multi-features in time series. The research object is a time series of real consumption records. We extract the consumption mark, frequency, and other characteristics. Furthermore, we use support vector machines (SVM), long short-term memory (LSTM), and other algorithms to forecast the user's consumption behaviour. We also used LSTM and SVM to perform multi-feature and multi-algorithm fusion.

Finally, the experimental results suggest that LSTM algorithms are more effective in prediction when the data is sparse. The SVM, on the other hand, is useful when data is ample. Furthermore, the LSTM-SVM fusion model provides advantages for both LSTM feature extraction and SVM classification. In most circumstances, LSTM-SVM performs the best in terms of prediction.Massive amounts of data are emerging as the Internet and information technologies advance at a rapid pace. As a result, data mining has emerged as one of today's most essential academic topics.

Time series analysis is a method for exploring all of the information contained in a time series, as well as observing, estimating, and studying the statistical regularity in the long-term change process in a set of real data [1]. Time series and data mining can be used to investigate changing rules of phenomena and forecast future behaviours.

Automated machine learning in practice: State of the art and latest findings, L. Tuggener, M. Amirian, K. Rombach, S. Lorwald, A. Varlet, C. Westermann, et al.,2019. The assumption that data-driven model building and decision-making may lead to increasing levels of automation and better informed judgements is a major motivation driving industry and society's digitisations. Machine learning is frequently used for creating such models from data. As a result, there is an ever- increasing demand for workers with the required skill set.

This requirement has given rise to a new research area, AutoML, which is focused with fully automatic machine learning model fitting. This paper provides an overview of the state of the art in AutoML, with a focus on practical applications in business, as well as recent benchmark results for the most prominent AutoML algorithms.

Many corporate and public organisations have recognised that data analysis is an effective tool for learning how to enhance their business model, decision- making, and even goods [1]. The design and training of a machine learning model is frequently a critical phase in the analytic process, and it comprises multiple difficult steps, most notably feature pre-processing, algorithm selection, hyper parameter tuning, and ensemble building.

Typically, extensive specialist knowledge is required to effectively complete all of these procedures [2]. The goal of automated machine learning (AutoML) is to develop strategies for creating suitable machine learning models with little (or no) human participation [3]. While the AutoML Problem can be stated in a variety of ways, we will focus on systems that handle the "Combined Algorithm Selection and Hyperparameter Optimisation" (CASH) problem [4]. A CASH problem solver seeks to select an algorithm from a list of possibilities and then tune it to provide the best validation performance across all algorithm combinations.

Predicting determinants for survival in breast cancer patients using machine learning techniques, M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon. 2019 [27]. Breast cancer is one of the most prevalent diseases among women worldwide. Many research have been conducted to predict survival indices; nevertheless, the majority of these analyses were carried out using simple statistical approaches. As an alternative, this study employed machine-learning approaches to create models for finding and visualising relevant prognostic indications of breast cancer survival rates.

This study employed a large hospital-based breast cancer dataset from the University Malaya Medical Centre in Kuala Lumpur, Malaysia (n = 8066) including diagnosis information from 1993 to 2016. The dataset had 23 predictor factors and one dependent variable that corresponded to the patients' survival status (alive or dead). Decision tree, random forest, neural networks, extreme boost, logistic regression, and support vector machine prediction models were used to identify significant predictive markers for breast cancer survival rates.

The information was then clustered based on the receptor status of breast cancer patients as determined by immunohistochemistry in order to do advanced modelling using random forest. Following that, the key variables were prioritised using random forest variable selection techniques. Finally, decision trees were developed and validated using survival analysis. In terms of model accuracy and calibration measure, all methods gave similar results, with decision tree having the lowest (accuracy = 79.8%) and random forest having the highest. This study found critical characteristics such as cancer stage classification, tumour size, number of total axillary lymph nodes excised, number of positive lymph nodes, primary therapy options, and diagnostic procedures.

Interestingly, the multiple machine-learning algorithms utilised in this study produced high accuracy, implying that these methods could be used as alternative prognostic tools in breast cancer survival studies, particularly in the Asian region. The significant prognostic factors impacting breast cancer survival rates discovered in this study, which were validated by survival curves, are useful and might be turned into medical decision support tools.

H. Tran, 2019, "A survey of machine learning and data mining techniques used in multimedia systems" [32]. Machine learning and data mining are computer science study areas that are rapidly developing due to breakthroughs in data analysis research, growth in the database business, and the accompanying market demand for methods capable of extracting valuable knowledge from enormous data stores. A great deal of research has been conducted in the multimedia field, focussing on various areas of data analytics, such as the capture, storage, indexing, mining, and retrieval of multimedia big data.

However, very few research works provide a comprehensive overview of the entire tree of approaches utilised in machine learning and data mining to solve research questions. In this survey study, we present a complete overview of cutting-edge methodologies, algorithms, machine learning, and data mining for multimedia systems. Query difficulty estimation predicts the search result performance of a particular query. It is a strong tool for retrieving multimedia, and it is gaining popularity.

There have been various strategies developed to evaluate query difficulty in textual information retrieval, but they cannot be used directly to picture search because they will result in poor performance. Currently, research on query difficulty estimation focusses on text-based enquiries, but there is a lack of study on image and video retrieval for multimedia queries. The widespread use of social media platforms such as Flickr, Facebook, and YouTube has significantly increased the Internet's multimedia database in recent years. These enhance database triggers may lead to the expansion of a huge number of multimedia study situations. The success of social media systems helps with content-based image retrieval.

S. D. Borde and K. R. Joshi, 2019. Improved signal detection algorithm for cognitive radio receivers utilising trained neural networks. Cognitive Radio has emerged as an important research topic in wireless communications in recent years. It has the potential to be extremely useful in dynamic spectrum management and interference identification.

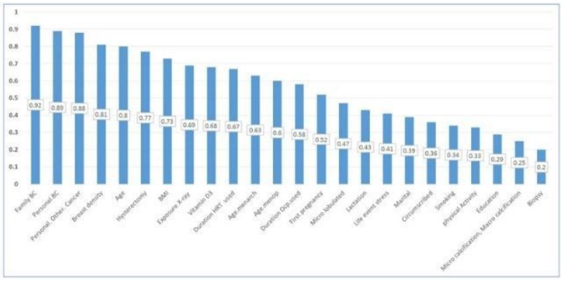
There are numerous spectrum-sensing algorithms proposed in the literature for cognitive radio, but all of them detect the presence or absence of the principal user in the allotted band and provide no information on the modulation scheme utilised. In some applications, such as cognitive radio receivers, it is important to recognise the modulation type of the signal in order to change the receiver parameters. Most modulated signals have the property of cyclo-stationarity, which can be utilised to correctly distinguish the principal user and the modulation type.

In this paper, we offer an upgraded signal recognition approach for cognitive radio receivers that utilises the modulated signal's cyclo-stationarity property to precisely determine the modulation type of the received signal using a trained neural network. The technique improves signal detection accuracy even in low SNR settings. The usage of a trained neural network increases its flexibility and extensibility for future applications.

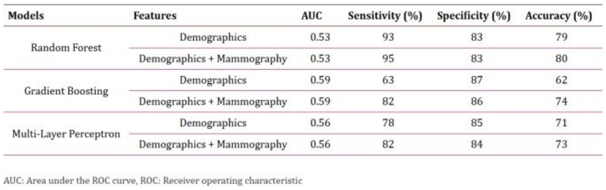
Performance of a decision tree C4.5 algorithm for evaluating student academic performance, Budiman, Haviluddin, Kridalaksana, Wati, and Purnawansyah, 2017 [39]. The new curriculum reform has resulted in considerable modifications to music courses in colleges and universities. As a result, student assessments in the classroom are evolving, and a more varied evaluation paradigm is progressively emerging. Numerous innovative and more effective teaching concepts and methods have been developed to revive the state through science and education. This disrupts the backward teaching pattern found in typical instructional exercises. As technology, science, and Internet technologies have evolved, online teaching has grown in importance in education.

. Music teachers at universities and colleges are always updating their teaching methods and utilising a variety of tactics to deliver in-depth training in the classroom. Many universities and colleges have widely implemented a web-based information educational administration management system to increase students' passion and involvement while also developing their musical creative talents. This work use the C4.5 algorithm to generate a decision tree model for developing a classroom teaching evaluation system to improve quality.

Using performance data from 125 teachers' music classroom lessons, the suggested algorithm assesses the model's correctness and practicability. Finally, it examines the decision-making traits that influence teacher assessment. The quality of classroom teaching is assessed, and some valuable suggestions are made based on the experimental results, which can aid college and university decision-making by inspiring teachers to enhance their classroom teaching quality.



The weight of the features in breast cancer prediction



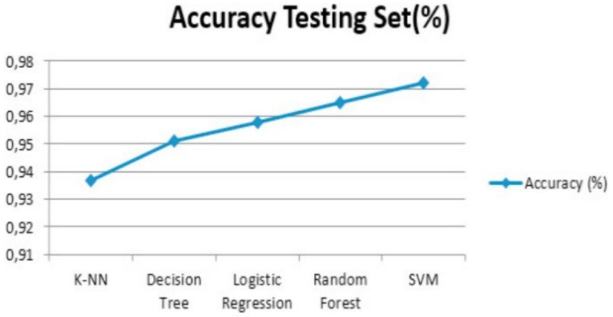
Models	Features	AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)
Random Forest	Demographics	0.53	93	83	79
	Demographics + Mammography	0.53	95	83	80
Gradient Boosting	Demographics	0.59	63	87	62
	Demographics + Mammography	0.59	82	86	74
Multi-Layer Perceptron	Demographics	0.56	78	85	71
	Demographics + Mammography	0.56	82	84	73

AUC: Area under the ROC curve, ROC: Receiver operating characteristic

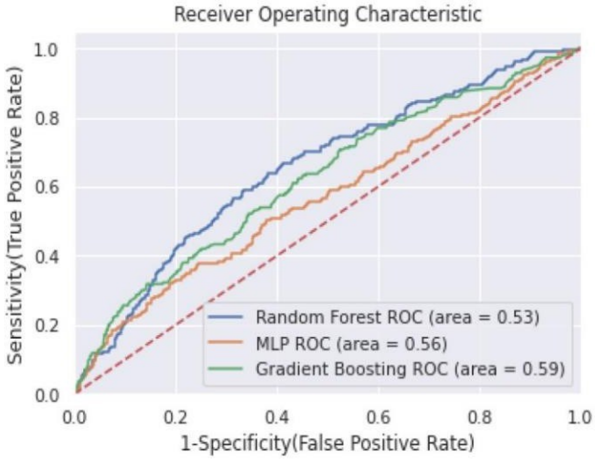
Performance comparison of the breast cancer prediction models

Algorithms	Accuracy Training Set (%)	Accuracy Testing Set (%)
SVM	98.4%	97.2%
Radom Forest	99.8%	96.5%
Logistic Regression	95.5%	95.8%
Decision Tree	98.8%	95.1%
K-NN	94.6%	93.7%

Accuracy percentage for breast cancer diagnostic dataset.



Comparative graph of different classifiers



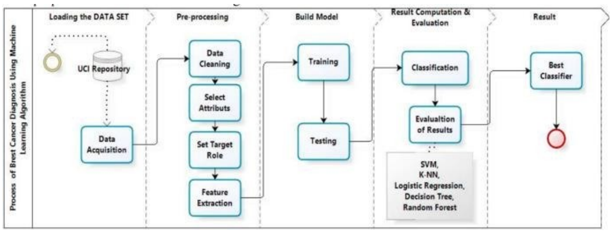
Receiver operating characteristic (ROC) curve of models

	Malignant	Benign	
SVM	201	11	Malignant
	1	356	Benign
Random Forest	196	16	Malignant
	7	350	Benign
Logistic Regression	201	11	Malignant
	5	352	Benign
C4.5	195	17	Malignant
	22	335	Benign
KNN	201	11	Malignant
	7	350	Benign

Confusion Matrix

Algorithms	Precision	Sensitivity	F-Measure	Class
SVM	0.98	0.94	0.96	Benign
	0.97	0.99	0.98	Malignant
Random Forests	0.96	0.94	0.95	Benign
	0.97	0.98	0.97	Malignant
Logistic Regression	0.98	0.91	0.94	Benign
	0.95	0.99	0.97	Malignant
Decision Tree	0.94	0.92	0.93	Benign
	0.96	0.97	0.96	Malignant
K-NN	0.92	0.91	0.91	Benign
	0.95	0.96	0.95	Malignant

Classifiers performances



Process Flow Diagram

3. LITERATURE SURVAY

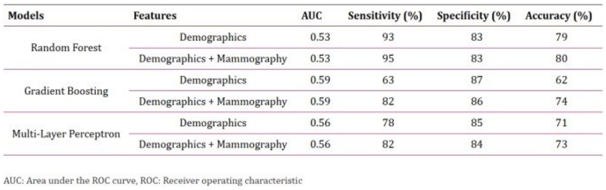
S.No	Title	Author	Journal Name & Year	Methodology Adapted	Key Findings	Gaps
1	Risk Factors and Preventions of Breast Cancer	Y.S. Sun, Z. Zhao, Z.N. Yang, F. Xu, H.J. Lu, Z.Y. Zhu, et al.	Inter National Journal of Biological Sciences, 2017.	Breast Cancer: Risk Factors and Prevention	Breast cancer aetiology, risk factors, and prevention	It may take photos directly through an X-ray-sensitive detector and analyse the digital data on a computer.
2	Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification	Y. Khourdifi and M. Bahaj	International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), 2018.	large data set, health system, prediction and classification.	Keywords help in result retrieval and provide a way to identify further relevant content.	Predict breast cancer, and with early identification and prevention, the risk of death is greatly reduced.
3	A Review of Breast Cancer Detection in Medical Images	Yao Lu, Jia-Yu Li, Yu-Ting Su and An-An Liu	IEEE Visual Communications and Image Processing (VCIP), 2018.	Image segmentation, registration, and fusion are all methods for detecting breast cancer.	Breast cancer, feature extraction and mammography.	Supervised learning, unsupervised learning, deep learning, CNN, and other methods for breast cancer detection are discussed.

4	Prediction of Upstaged Ductal Carcinoma In Situ Using Forced Labeling and Domain Adaptation	RuiHou, Maciej A. Mazurowski, Lars J. Grimm, Jeffrey R. Marks, Lorraine M. King, Carlo C. Maley, Eun-Sil Shelley Hwang and Joseph Y. Lo	IEEE Transactions on Biomedical Engineering (Volume: 67, Issue: 6, June 2020)	To enhance the prediction of pure ductal carcinoma in situ (DCIS) (negative) versus upstaged DCIS (positive) cases.	Breast cancer, domain Adaptation, Forced labeling	We improved the prediction performance of DCIS upstaging by integrating two related pathology classes in different training periods.
5	Efficacy of the Multidisciplinary Tumor Board Conference in Gynecologic Oncology: A Prospective Study	Banghyun Lee, Kidong Kim, Jin Young Choi, dong hoonSuh, Jae Hong No, Ho-Young Lee, Keun Yong Eom, Haeryoung Kim, Hak Jong Lee, Yong Beom Kim	Medicine (Baltimore), 2017	Radiation oncologists, urologists, general surgeons, and orthopedists	Conference, diagnosis, diagnostic techniques and procedures, oncology, therapeutics	The alterations in diagnosis, diagnostic work-up, and radiological results represented the gynaecologic oncology MTBC.
6	A Comprehensive Survey of Data Mining	Manoj Kumar Gupta, Pravin Chandra.	International Journal of Information Technology, 2020	Banking, retail, medical, insurance, bioinformatics	Data mining, patterns, capabilities	A systematic and comprehensive review of various data mining tasks and methodologies
7	Breast cancer detection by leveraging Machine Learning	Anji Reddy Vaka, BadalSoni, Sudheer Reddy K.	ICT Express Volume 6, Issue 4, December 2020	Data set, Data augmentation, Pre-processing, Feature extraction, Histo-sigmoid fuzzy clustering	Machine Learning, Classification, Breast cancer, Deep learning	The proposed method is based on Support value on a deep neural network. To meet the better performance, efficiency, and quality of images

8	Performance evaluation of machine learning for breast cancer diagnosis: A case study	MostafaShanbehzadeh, HadiKazemi-Arpanahi, Mohammad BolbolianGhalibaf, AzamOrooji	Informatics in Medicine Unlocked. Volume 31, 2022	Data collection, Data pre-processing, Identification of risk factors influencing the BC diagnosis, Predictive model development	Machine learning, Artificial intelligence, Data mining, Breast neoplasms	The generated models may effectively categorise individuals who are at high risk for cancer, and can be used as a screening tool for the early Breast cancer detection
9	Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review	S. Eltalhi and H. Kutrani	IOSR Journal of Dental and Medical Sciences (IOSR-JDMS) Volume 18, Issue 4 Ser. 20 (April. 2019)	Data set, Data Collection Surgery, Attributes Rediotherapy, Chemotherapy, Hormone therapy, Biological therapy	Breast cancer, machine learning, data mining, classification algorithms, clustering algorithms	Breast cancer prediction models include Decision Trees, Naïve Bayes, and Artificial Neural Networks.
10	Research on machine learning algorithms and feature extraction for time series	Lei Li, Yabin Wu, YihangOu, Qi Li, Yanquan Zhou, Daoxin Chen	IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017	Dataset, Time series analysis, Prediction algorithms, Feature extraction, Prediction.	Support vector machines, Time series analysis, Prediction algorithms, Feature extraction, Machine learning algorithms, Predictive models, Data mining	Multi-feature and multi-algorithm fusion using LSTM and SVM were implemented.
11	Automated Machine Learning in Practice: State of the Art and Recent Results	Lukas Tuggener, MohammadrezaAmirian, Katharina Rombach, Stefan Lorwald, Anastasia Varlet, Christian Westermann, ThiloStadelmann	Swiss Conference on Data Science (SDS), 2019.	Data set, Data collection, pre-processing Analysis, Feature extraction,	Optimization, Machine learning, Pipelines, Data models, Feature extraction, Buildings, Machine learning algorithms	AutoML is a research field that focusses on entirely autonomously fitting machine learning models.
12	Predicting	MoganaDarshi	BMC Medical	Data collection,	Machine	Machine

	factors for survival of breast cancer patients using machine learning techniques	niGanggayah, NurAishahTaib, Yip Cheng Har, PietroLio&SarinderKaurDhillon	Informatics and Decision Making, 2019	Model evaluation, Random forest advanced modelling, Variable selection, Survival analysis	learning, decision tree, Breast cancer	learning approaches are used to develop models for discovering and visualising relevant prognostic factors of breast cancer survival rate.
13	Survey of Machine Learning and Data Mining Techniques used in Multimedia System	Hieu Tran	A Preprint, 2019	Image mining, Multimedia Contents, Spatiotemporal Segmentation, Feature Extraction, Evaluation of Result.	Big Data Analytics, Multimedia analysis, Multimedia databases, Indexing, Retrieval	Multimedia mining fundamental principles, important properties, architectures, models, and applications
14	Enhanced signal detection algorithm using trained neural network for cognitive radio receiver	SheetalBorde, Kalyani Rajeev Joshi	International Journal of Electrical and Computer Engineering (IJECE), February 2019.	Enhanced Signal Detection (ESD) algorithm, Cyclic frequency Domain Profile (CDP), Back Propagation Algorithm.	Artificial neural network, Cognitive radio, Cyclostationarity , Enhanced signal detection, Spectrum sensing	An upgraded signal detection technique for a cognitive radio receiver that uses the cyclostationarit y property of the modulated signal to precisely detect

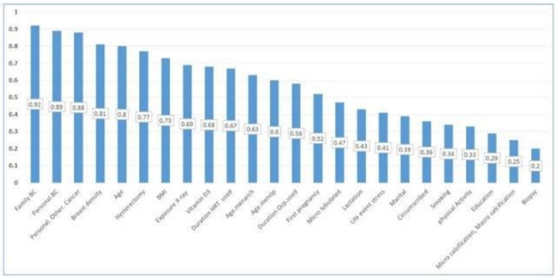
4. RESULTS



Models	Features	AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)
Random Forest	Demographics	0.53	93	83	79
	Demographics + Mammography	0.53	95	83	80
Gradient Boosting	Demographics	0.59	63	87	62
	Demographics + Mammography	0.59	82	86	74
Multi-Layer Perceptron	Demographics	0.56	78	85	71
	Demographics + Mammography	0.56	82	84	73

AUC: Area under the ROC curve, ROC: Receiver operating characteristic

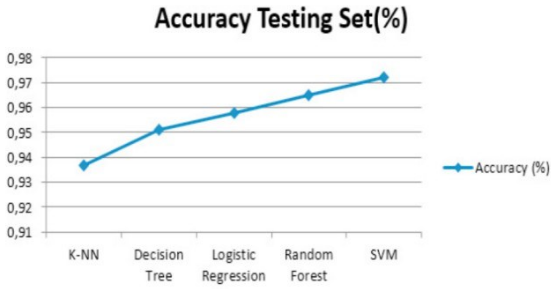
The weight of the features in breast cancer prediction



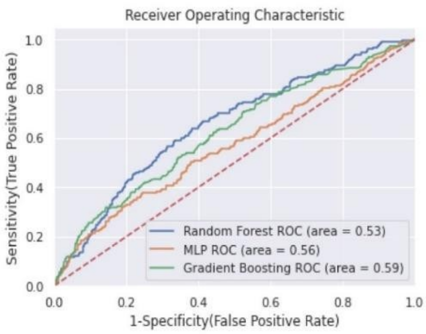
Performance comparison of the breast cancer prediction models

Algorithms	Accuracy Training Set (%)	Accuracy Testing Set (%)
SVM	98.4%	97.2%
Radom Forest	99.8%	96.5%
Logistic Regression	95.5%	95.8%
Decision Tree	98.8%	95.1%
K-NN	94.6%	93.7%

Accuracy percentage for breast cancer diagnostic dataset.



Comparative graph of different classifiers



Receiver operating characteristic (ROC) curve of models

	Malignant	Benign	
SVM	201	11	Malignant
	1	356	Benign
Random Forest	196	16	Malignant
	7	350	Benign
Logistic Regression	201	11	Malignant
	5	352	Benign
C4.5	195	17	Malignant
	22	335	Benign
KNN	201	11	Malignant
	7	350	Benign

Confusion Matrix Classifiers performances

5. CONCLUSION

In this article, we looked at numerous machine learning, deep learning, and data mining strategies for predicting breast cancer. Our goal is to find the best algorithm for correctly predicting breast cancer incidence rates. The major purpose of this review is to highlight all prior research on machine learning algorithms used to detect breast cancer. This article contains all of the knowledge that novices need to examine machine learning algorithms and develop a good foundation in deep learning. The review of this article begins with the many different types of breast cancer; fourteen research papers were assessed to provide an overview of the basic types, symptoms, and causes of breast cancer. Following that, an overview of the primary machine learning techniques, ensemble approaches, and deep learning techniques was provided, all of which are extremely complex algorithms used to predict breast cancer.Some concerns must be addressed in future development. Researchers can employ data augmentation methodologies to deal with the issue of limited dataset availability. Researchers should consider the issue of positive and negative data inequality, as this might lead to bias in positive or negative predictions. Another significant issue that needed to be addressed was the disparity in the number of breast cancer photos vs impacted patches for accurate diagnosis and prognosis of breast cancer.

REFERENCES

1.Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, et al., "Risk factors and preventions of breast cancer", Int. J. Biol. Sci., vol. 13, pp. 1387, 2017.

2.Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification", Proc. Int. Conf. Electron. Control Optim. Comput. Sci. (ICECOCS), pp. 1-5, Dec. 2018.

3.Y. Lu, J.-Y. Li, Y.-T. Su and A.-A. Liu, "A review of breast cancer detection in medical images", Proc. IEEE Vis. Commun. Image Process. (VCIP), pp. 1-4, Dec. 2018.

4.F. K. Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier", Proc. 13th Int. Conf. Intellient Syst. Design Appl., pp. 121-125, Dec. 2013.

5.R. Hou, M. A. Mazurowski, L. J. Grimm, J. R. Marks, L. M. King, C. C. Maley, et al., " Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation ", IEEE Trans. Biomed. Eng., vol. 67, no. 6, pp. 1565-1572, Jun. 2020.

6. A. R. Chaudhury, R. Iyer, K. K. Iychettira and A. Sreedevi, "Diagnosis of invasive ductal carcinoma using image processing techniques", Proc. Int. Conf. Image Inf. Process., pp. 1-6, Nov. 2011.

7. S. Pervez and H. Khan, " Infiltrating ductal carcinoma breast with central necrosis closely mimicking ductal carcinoma in situ (comedo type): A case series ", J. Med. Case Rep., vol. 1, no. 1, pp. 83, Dec. 2007.

8. D. L. Page, W. D. Dupont, L. W. Rogers and M. Landenberger, "Intraductal carcinoma of the breast: Follow-up after biopsy only", Cancers, vol. 49, no. 4, pp. 751-758, 1982.

9. A. B. Tuck, F. P. O'Malley, H. Singhal and K. S. Tonkin, "Osteopontin and p53 expression are associated with tumor progression in a case of synchronous bilateral invasive mammary carcinomas", Arch. Pathol. Lab. Med., vol. 121, no. 6, pp. 578, 1997.

10. B. Lee, K. Kim, J. Y. Choi, D. H. Suh, J. H. No, H.-Y. Lee, et al., "Efficacy of the multidisciplinary tumor board conference in gynecologic oncology: A prospective study", Medicine, vol. 96, no. 48, pp. e8089, Dec. 2017.

11. S. Masciari, N. Larsson, J. Senz, N. Boyd, P. Kaurah, M. J. Kandel, et al., "Germline E-cadherin mutations in familial lobular breast cancer", J. Med. Genet., vol. 44, no. 11, pp. 726-731, Aug. 2007.

12. A. Memis, N. Ozdemir, M. Parildar, E. E. Ustun and Y. Erhan, "Mucinous (colloid) breast cancer: Mammographic and US features with histologic correlation", Eur. J. Radiol., vol. 35, no. 1, pp. 39-43, Jul. 2000.

13. A. Gradilone, G. Naso, C. Raimondi, E. Cortesi, O. Gandini, B. Vincenzi, et al., "Circulating tumor cells (CTCs) in metastatic breast cancer (MBC): Prognosis drug resistance and phenotypic characterization", Ann. Oncol., vol. 22, no. 1, pp. 86-92, Jan. 2011.

14. F. M. Robertson, M. Bondy, W. Yang, H. Yamauchi, S. Wiggins, S. Kamrudin, et al., "Inflammatory breast cancer: The disease the biology the treatment", CA Cancer J. Clin., vol. 60, no. 6, pp. 351-375, 2010.

15. M. K. Gupta and P. Chandra, "A comprehensive survey of data mining", Int. J. Inf. Technol., pp. 1-15, Feb. 2020.

16. D. Delen, "Analysis of cancer data: A data mining approach", Expert Syst., vol. 26, no. 1, pp. 100-112, Feb. 2009.

17. M. Shahbaz, S. Faruq, M. Shaheen and S. A. Masood, "Cancer diagnosis using data mining technology", Life Sci. J., vol. 9, no. 1, pp. 308-313, 2012.

18. A. Reddy, B. Soni and S. Reddy, "Breast cancer detection by leveraging machine learning", ICT Express, 2020.

19. Z. Salod and Y. Singh, "Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol", J. Public Health Res., vol. 8, no. 3, pp. 1677, Dec. 2019.

20. S. Eltalhi and H. Kutrani, "Breast cancer diagnosis and prediction using machine learning and data mining

21. M. Rana, P. Chandorkar, A. Dsouza and N. Kazi, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", Int. J. Res. Eng. Technol., vol. 4, no. 4, pp. 1163-2319, 2015.
22. P. Israni, "Breast cancer diagnosis (BCD) model using machine learning", Int. J. Innov. Technol. Exploring Eng., vol. 8, no. 10, pp. 4456-4463, Aug.
2019.
23. L. G. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi and A. R. Razavi, "Using three machine learning techniques for predicting breast cancer recurrence", J. Health Med. Inform., vol. 4, no. 124, pp. 3, 2013.
24. E. A. Bayrak, P. Kirci and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis", Proc. Sci. Meeting Elect.-Electron. Biomed. Eng. Comput. Sci. (EBBT), pp. 1-3, Apr. 2019.
25. M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, et al., "A new nested ensemble technique for automated diagnosis of breast cancer", Pattern Recognit. Lett., vol. 132, pp. 123-131, Apr. 2020.
26. D. A. Omondiagbe, S. Veeramani and A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis", IOP Conf. Ser. Mater. Sci. Eng., vol. 495, Jun. 2019.
27. S. N. Singh and S. Thakral, "Using data mining tools for breast cancer prediction and analysis", Proc. 4th Int. Conf. Comput. Commun. Automat. (ICCCA), pp. 1-4, Dec. 2018.
28. M. J. Zaki and W. Meira, Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Cambridge, U.K.:Cambridge Univ. Press, 2019.
29. S. Aruna and S. Rajagopalan, "A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer", Int. J. Comput. Appl., vol. 31, no. 8, pp. 1-7, 2011.
30. 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, no. 4, pp. 1476-1482, Mar. 2014.
31. C. Shah and A. G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction", Proc. 4th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), pp. 1-4, Jul. 2013.
32. L. S. Jamil, "Data analysis based on data mining algorithms using weka workbench", Int. J. Eng. Sci. Res. Technol., vol. 5, no. 8, pp. 262-267, 2016.
33. G. R. Kumar, G. Ramachandra and K. Nagamani, "An efficient prediction of breast cancer data using data mining techniques", Int. J. Innov. Eng. Technol., vol. 2, no. 4, pp. 139, 2013.
34. A. A. Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer

detection", Int. J. Mach. Learn. Comput., vol. 9, no. 3,
pp. 248-254, Jun. 2019.
35. A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques", Proc. SIAM Int. Conf. Data Mining, vol. 58, pp. 10- 110, 2006.
36. J. Talukdar and S. K. Kalita, "Detection of breast cancer using data mining tool (weka)", Int. J. Sci. Eng. Res., vol. 6, no. 11, pp. 1124, 2015.
37. B. Padmapriya and T. Velmurugan, "Classification algorithm-based analysis of breast cancer data", Int. J. Data Mining Techn. Appl., vol. 5, no. 1, pp. 43-49, Jun. 2016.
38. K. Williams, P. A. Idowu, J. A. Balogun and A. I. Oluwaranti, "Breast cancer risk prediction using data mining classification techniques", Trans. Netw. Commun., vol. 3, no. 2, pp. 1, Apr. 2015.
39. S. Bharati, M. A. Rahman and P. Podder, "Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA", Proc. 4th Int. Conf. Electr. Eng. Inf. Commun. Technol. (iCEEiCT), pp. 581-584, Sep. 2018.
40. P. Mekha and N. Teeyasuksaet, "Deep learning algorithms for predicting breast cancer based on tumor cells", Proc. Joint Int. Conf. Digit. Arts Media Technol. With ECTI Northern Sect. Conf. Electr. Electron. Comput. Telecommun. Eng. (ECTI DAMT-NCON),
pp. 343-346, Jan. 2019.
41.