Enhanced machine learning models for predicting breast cancer: healthcare system

Dilshad Fadhil Mawlood[1*], *Dona* A. Franci[2], *Darun* Mudhafar Hamad[3], Shahab Wahab Kareem[4]

[1] College of Engineering, Department of Information System Engineering, Erbil Technical Engineering College, Erbil Polytechnic University, Erbil, Iraq
[2,3,4] College of Science and Engineering, Department of Information Technology, Catholic University in Erbil, Erbil, Iraq.

Abstract Currently, breast cancer is a popular illness that can lead to many consequences, with the most severe outcome being death rates. Therefore, there is a pressing requirement for a diagnostic tool that can aid healthcare professionals in early detection of the illness and provide required lifestyle modifications to prevent its development the possibility of developing cancer at a young age has also been significantly enhanced by environmental alterations in our daily existence. This analysis aimed to accurately classify features into either malignant or benign classes. The suggested methodologies and classifying systems were applied to the Wisconsin Diagnostic Breast Cancer (WDBC) and Breast Cancer Coimbra Dataset (BCCD) datasets. Conventional performance measures, such as (KNN, SVM, ensemble classifier (EC), and logistic regression (LR)) methods, were utilized to evaluate the efficacy and time of training for each classifier. The diagnostic power of the models was enhanced by our DET (Diagnostic Enhancement Technique). Specifically, the polynomial SVM achieved an accuracy of 98.3%, LR (Logistic Regression) reached 97.04%, KNN (K-Nearest Neighbors) achieved 96.3%, and EC (Ensemble Classifier) achieved 96.6% accuracy with the dataset is called WDBC. In addition, in this study, there's just make a comparative analysis of the findings in relation to the accuracy of the outcomes of prior research. The implementation process and results can assist clinicians in adopting an efficient prototype for functional comprehension and forecast of breast cancer (BC) tumours.

1 Introduction

Various types of habitats have been identified to impact different cancer risks, such as exposure to UV radiation, pollution, and habitat fragmentation. The impact of these diverse cancer risk factors varies depending on the specific environment, including UV radiation, pollutants, and habitat fragmentation. Machine learning has been utilized not only to classify different habitat types in various biomes but also to identify human activities that lead to detrimental changes in the ecosystem, such as fragmentation and pollution from oil spills.

*Corresponding author: eng.dlshadfm@gmail.com

The emerging field of machine learning in habitat mapping will offer valuable insights into how cancer affects animals and their habitats. The acceptance of machine learning in the medical and healthcare sectors has increased due to its ability to rapidly and accurately analyze vast amounts of data, leading to improved disease prediction. Machine learning can be a valuable tool for healthcare professionals in identifying risk factors, making diagnoses, and predicting disease progression by extracting patterns from clinical data, including genomic information, patient records, and images. These models can be customized to predict the likelihood of a patient developing a specific disease, identify risk factors, diagnose illnesses, predict disease progression, and assist in selecting the most appropriate treatment for individual patients. Four distinct prediction models were created using four machine learning algorithms (KNN, SVM, ensemble classifier (EC), and logistic regression (LR)) to analyze a large dataset of tumor characteristics for the diagnosis of breast cancer (BC). The goal became to investigate a precise and effective prediction version for tumor categorization via the utilization of information mining strategies. There's an suggests using a four-layered method called sizeable records exploratory strategies (DET) to investigate the Breast Cancer Coimbra Dataset (BCCD) and Wisconsin Diagnostic Breast Cancer (WDBC). These approaches involve examining the distribution of features, eliminating irrelevant ones, and generating a hyperparameter for practical analysis. These strategies allowed the enhancement of accuracy and diagnostic efficiency in machine learning predictive models. The structure of (WDBC) and (BCCD) is combines DET and predictive models to investigate about the how of the method for nursing analysis of breast most cancers (BC). The tumour characteristics can be expounded to a level of worst terrific detail and this may lead to a lot of repeats. These characteristics lead to tedious results due to long computation times that are associated with the frameworks. Such a perspective of time performance will enable the models to sift through large datasets and subsequently isolate records containing vital information from relationships and discard irrelevant features. A high number of deaths from breast cancer was diagnosed with the help of the models created, while also bringing about the shortest time for computational calculations. Overall, these outcomes with the models will give a facts analyst using a mechanism to implement a state-of-the-art system learning version in distinguishing breast cancer. The next parts of the article are structured in the following manner: Section two elaborates on the article reviews, Section three presents the proposed approach, Section four focuses on evaluating the results, Section five discusses the findings, and Section six concludes the study.

2 Related works

Right now, breast [2] cancer accounts for around 2,350,500 novel cases and 663,976 novel passings in 2020, and is the moment most common executioner illness in ladies universally moment as it were to lung cancer. Collectively, over the Joined together States, a add up to of 271,550 unused rates of (BC) were recognized in 2021, with 42,500 detailed cases of passing among females [2]. Breast cancer (BC) is cancer that starts within the breast tissues, beginning most frequently within the internal lining of either the drain creating ducts or the lobules that produce drain. Cancer may be a alter within the ordinary condition of cells whereby they change into cancer cells due to changes within the hereditary cosmetics more often than not RNA and DNA. This could be because of the arbitrary hereditary float which increments entropy or it may be due to other causes. A few cases of the perilous specialists grasping: electromagnetic radiation which is within the shape of bright beams, microwaves, X-rays and gamma beams, etc. The sources of potential hurt incorporate atomic radiation, microorganisms such as microbes, infections, and organisms, parasites, airborne poisons, tall temperatures, sullied nourishment and water, as well as free radicals, mechanical damage at the cellular level, DNA and RNA advancement, and maturing [3]. Generous and cancerous

tumors are two isolated classifications. In spite of the fact that generous tumors are not as perilous as cancerous ones, they can in any case raise the probability of getting breast cancer. On the other hand, dangerous tumors are more concerning and cancerous. A look at carried out on breast cancer distinguishing proof decided that 21% of ladies capitulated to harmful tumors [4]. This watch makes a forte of the distinguishing proof and lesson of tumors, that's directly an exceptional subject matter within the teach of biomedicine. The analysts are utilizing system studying (ML) and truths mining (DM) era to estimate breast cancer [5]. Applying classifier-based completely expectation models interior the spaces of information mining (DM) and contraption examining (ML) may moreover enormously decrease demonstrative mistakes and upgrade the execution of cancer guess. Information mining (DM) could be a complex total of numerous techniques utilized to find stowed away data and data from huge databases that are challenging to require a see at straightforwardly. It has been significantly utilized interior the presentation of prognostic structures for various infections, which incorporates heart illness [6], lung cancer [7], and thyroid most cancers [8]. Computer-aided structures that coordinated actualities mining (DM) and gadget acing (ML) techniques have been utilized to analyze breast cancer, which incorporate the utilization of fluffy hereditary qualities [10] [9]. The studyâ€™s discoveries viably categorize the characteristics into two tumor classes the utilization of a classifier and make forecasts roughly fate tumors basically based on past records.

An investigation carried out certainly set up that the utilization of gadget getting to know classifiers for ensuing organize of breast most cancers adjusted off most cancersâ€™ expectation in early arrange unquestionably up the extent of survival and makes a difference to control the dissemination of most cancersâ€™cells around the body [11]. An occurrence may be a look at that utilized the back vector equipped up methodology (SVM) based procedure in determination of breast cancer in expansion to gotten promising comes about on expectation [12]. Moreover, use Furey et. Another consider, [13], a combination of consistency-based highlight determination strategy and Bolster Vector Machines (SVM) with direct part was utilized to classify cancer tissue and accomplished an exactness of 93. 4%. Along these lines, Zheng et. In fact, Jeong et al. (2014) progressed on this investigate by setting up a K-SVM combined form for the categorization of Wisconsin Demonstrative Breast Cancer (WDBC) set. Their form achieved an precision of 97% [14]. In the interim, some analysts centered on creating elective classifiers. For case, Seddik et. (2015) brought a procedure that connected tumor components in a double calculated adaptation to hit upon breast most cancers utilizing WDBC data. This approach yielded favorable results [15]. In a comparative way, Mert and colleagues enlisted a okay-nearest neighbor (KNN) classifier to estimate breast most cancers. They finished this by utilizing developing a include markdown method the utilization of fair-minded perspective investigation. The framework scattered the capabilities by means of diminishing one characteristic (1C) and the utilization of 30 capabilities. It at that point computed the generally execution and wrapped up an precision of 90.1% [16]. In expansion to the advantageous exactnesses accomplished with different classifiers and calculations, it is worth noticing that the previously mentioned investigate has not considered the utilization of information exploratory approaches. These procedures upgrade the vigor of information mining strategies, coming about in more proficient execution. The need of these crucial techniques ruins the precision of ML classifiers in a few inquire about [16â€“ 19]. Be that as it may, within the oneâ€™s examinations, the perplexity frameworks incorrectly labeled the dangerous and kind preparing since they erroneously expected the veritable loathsome and wrong destitute matrices. A encourage blemish got to be analyzed interior the previously mentioned ponders that utilized benchmarks to evaluate the highlight preparing beside nonlinear category. In any case, the execution time of the form significantly raises since the wide assortment of capacities climbs [20]. Progressed innovation, in conjunction with a run of information science techniques,

help within the gathering and investigation of cancer-related information to figure this life-threatening sickness. Machine learning calculations have viably been utilized within the consider of cancer-related information, nearby other information science innovations. An illustrative occasion of investigate [21] was embraced to substantiate the upgrade of symptomatic exactness by using machine learning strategies. An master doctor accomplished a conclusion precision of 79.97%. All things considered, machine learning accomplished an exactness rate of 91.1% in its forecasts. Over the past few decades, the utilization of machine learning within the therapeutic space has continuously developed. Be that as it may, the basic components for investigation incorporate the information collected from the sufferers and the appraisal executed through the clinical master.Machine learning classifiers have successfully minimized human botches and brought actuate examination of logical data with progressed comprehensiveness [22]. Furey et. In 2000, the creators Zhang Wei and Kong Lingjie utilized Back Vector with direct bit for the distinguishing proof of cancerous tissue and the comes about were palatable precise are [2]. Additionally, Polat along with others have famous that the social judgment definition holds that, fair as the ways in which individuals think, do, and feel are impacted by the discrete response propensities that underlie them, so as well is their social conduct influenced by these inclinations. Direct relapse vector machine (SVM) utilized in D.M Gunawardane-Information Sciences (2007)) to foresee breast cancer and the exactness was 98%. failure to perform quicker and diminish mistakes by getting freed of insignificant property which seem reach an astonishing 53% precision. The creators considered utilizing slightest square SVM for show preparing with direct conditions, based on a earlier suggestion [23]. Even so, his approach fizzled to supply the method of selecting highlights. The creator [24] created a disseminated database that joins numerous advances and underpins mulu-active highlights. In 2010, Prasad and Jain et. [25] presented a heuristic method for selecting a subset of highlights to prepare the SVM classifier. The BC information is classified into two particular classes with an precision of 91.7%. Indeed so, the creator can essentially upgrade this precision by utilizing the include end strategy to dispense with the nearness of loud information. In a comparative way, Zheng et. (2014) put out a cross breed demonstrate that consolidates K-means and SVM classifiers. The essential point of this demonstrate was to utilize the include choice and extraction method to analyze the characteristics of tumors within the Wisconsin Demonstrative Breast Cancer (WDBC) dataset. K-nearest neighbor of classification sort known as K-Means was utilized to distinguish between benign and malignant tumor designs. As modern preparing designs, the computed designs are created and after that utilized to upgrade the SVMâ€™s show. Advance, the yield produced is utilized to foresee end of the cancer rates utilizing Back Vector Machine (SVM) calculation. The work that was done taking after their demonstrate of coordinates frameworks come about into having a fair accuracy of 97%. Be that as it may, the strategy of records exploratory that's so crucial for the instructing of statistics elucidation has not been agreed prevailing accentuation to instructing the prescribed kind [15].Some of the other methods adopted are: Support Vector Machines (SVM) Support Vector Machines (SVM) in conjunction with other techniques Neural Networks K-Nearest Neighbour (New) and more. With the help of Most Important Parameters and Wisconsin Original Breast Cancer (WOBC), as well as Wisconsin Digital Breast Cancer (WDBC) datasets, the study of (Aydin & Yalcin, 2013) used Logistic Regression (LR). They managed to implement the version that achieved a best accuracy of 97. 21% for the TEMP dataset and 92 for the WOBC dataset. Optimized characteristic units [26] is 88% for WDBC dataset. In a similar context, Seddik et al. Breast cancer was diagnosed the usage of variables that represent tumour imaging properties, by Bernard et.al in 2015 binary logistic mode. The proposed model as it should be classifying the WDBC statistics into malignant and benign classes, with an average classification accuracy of 98.01%. The regression version diagnosed region, texture, concavity, and symmetry as statistically huge variables inside the WDBC dataset [15]. Other

systematic reviews comparing existing works have found even more articles that employed the SVM model in breast cancer classification. Yet, the usage of alternative models was sparse in only a selected number of studies. There was a technique; It was posed by A. Mert et. These characteristics were identified using Independent Component Analysis (ICA) in the year 2015 with the aim of reducing the number of characteristics with an intention of making a prediction about the Breast Cancer. Thus, the attribute ranking turned into applied to the WDBC data set and efficiently classified utilising both 1C or all 30 features with the k-nearest neighbour (KNN) classifier. The performance was measured using several matrices as presented below and had a thirty-one-point one percent accuracy level. 1% [16]. Subsequently, Rajaguru et. Extensions far from this study were seen in; the authors made improvements on it by employing The KNN and choice tree (DT) gadget studying algorithms in an effort to correctly categorise the WDBC dimensions inside the breast cancer prediction problem. The study employed a conventional principal component analysis (PCA) technique to extract features for categorization. The results indicated that KNN performed better than DT [18]. Yang and Xu et. (2019) did a study where KNN attained an accuracy of 96.4% using the same feature selection approach, namely Principal Component Analysis (PCA) [27]. Lately, my work has centered on evaluating the efficiency of K nearest friends (KNN) set of rules through varying the k values and exploring exceptional distance functions. The goal is to determine the usefulness of KNN on two distinct breast cancer datasets. The experiment has 3 wonderful types: KNN with Chi-square-based totally features, KNN with linear SVM and KNN without characteristic choice, and KNN with Chi-square-primarily based capabilities. The consequences confirmed that the 1/3 method, Chi-rectangular-based characteristic choice, achieved the maximum accuracy on both datasets while the use of Manhattan or Canberra distance functions [19]. Regarding the fourth prediction model, the ensemble classifier (EC) with the voting approach, there is restrained research on its application in breast cancer prediction. For instance, in their look at, M. Abdar et. (2020) introduced an ensemble method the use of a vote casting classifier to appropriately discover benign tumors from malignant breast cancer. A two-layer vote casting classifier turned into applied the use of two or three distinct system learning algorithms. The voting procedures revealed the satisfactory performance of the straightforward classification algorithm [5]. A recent publication in Nature Cancer introduced a methodology for categorizing cancer cells as either normal or malignant tissues [28]. Several researches have employed the SVM classifier for breast cancer prediction, with best a handful of them making use of a single classifier of their exams. Nevertheless, there's a persisted need to investigate a greener classifier for predicting breast most cancers using greater powerful techniques [14,15]. This examine hired 4 awesome prediction fashions utilizing strong exploratory strategies in facts mining. For the motive of diagnosing breast cancer (BC). Figure 1, 2, and 3 show three distinct prediction models had been developed the usage of three machine learning algorithms (KNN, SVM, and logistic regression (LR)) to handle a huge amount of tumor traits if you want to retrieve vital information for diagnosing breast cancer (BC).

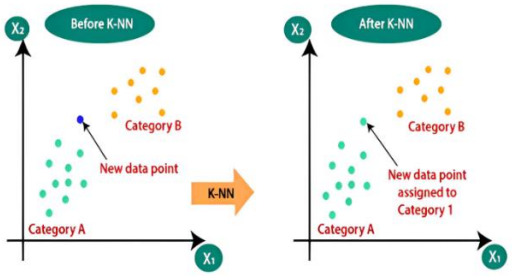

Fig. 1. The K-Nearest-Neighbor-Algorithm-for-Machine-Learning [46]

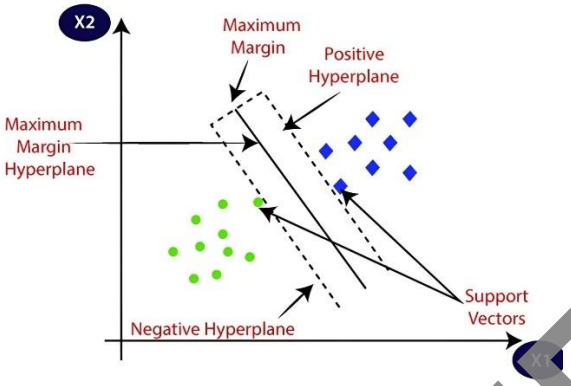

Fig. 2. The Support-Vector-Machine-Algorithm [46]

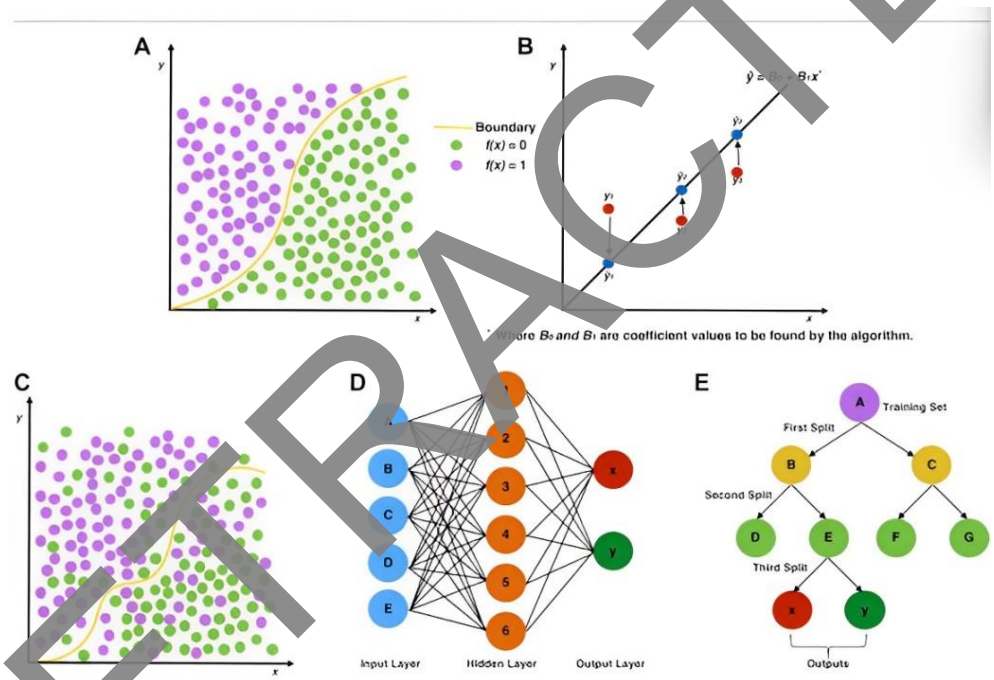

Fig. 3. The Machine Learning Versus Logistic Regression Methods [47]

A- linear regression, B- support vector machine, C- artificial neural network, D- decision tree, E- approaches

3 Methodology

Most chronic diseases may be predicted with a reasonable degree of mastery with the help of machine learning systems. The machine learning model's input is structured data, that is, data with a specific format and architecture. This system is used by end users such as a patient or any end user of the system in general. All these various options are going to be entered into this system by the user. An example of an input type will be numerical data that the patient will input in the specified medical diagnostic record. The quantitative data from these cells will be given to the machine learning algorithm for disease diagnosis. They are then used to identify the accuracy to be used by diagnosing the problem appropriately. This technology is going to use the symptoms that are experienced by the patient to predict whether the tumour is of the malignancy type or the benign type. This approach incorporates the element of Machine Learning Technology in order to predict diseases from symptoms. Naïve Bayes is used for disease prediction, ADA boost for classification, logistics regression for feature extraction, as well as for division of /datasets using SVM. The last output of the system is the breast cancer prediction generated from the top performing algorithm among those suggested here.
This part will cover the proposed technique, which encompasses model architecture, data information, Machine learning models and the criteria used to evaluate them.

3.1 Creative Framework

In this study, via address the following issues related to the breast cancer dataset; which can be freely downloaded from the reference [16-19].
In the specific context of breast cancer, it might be useful to know when and how DET should be combined with the rest of the aforementioned prediction models from a specifically efficient angle. How might the characteristics of breast cancer aid in enhancing the accuracy and scalability of machine learning models in detecting cancer?
In order to address these issues, a possible resolution is presented, as depicted in Figure 4. This solution consists of nine distinct and consequential actions. The key constituents of this approach are listed below:

A. As for the database resources, the WDBC and the BCCD are collected from the ML database.
B. The onset of the data preparation involves the accomplishment of certain fundamental cycles for each discrete data.
C. To classify the data in the first dataset, the WDBC dataset, as either malignant or benign and in the second dataset, BCCD dataset as either present or absent.
D. Determine whether two attributes are good, bad, or random (independent of one another) based on values versus values correlation test.
E. To achieve better results, one should note less important factors and then retain such recursive characteristics to make outcomes more efficient.
F. As a result of exploratory data analysis, split the gathered data into different training data and testing data.
G. Use of ids in the datasets as an input for four models, namely the SVM, LR, KNN, and EC.
H. Finally, perform an evaluation of the values and compare them to each model in relation to studies previously conducted.

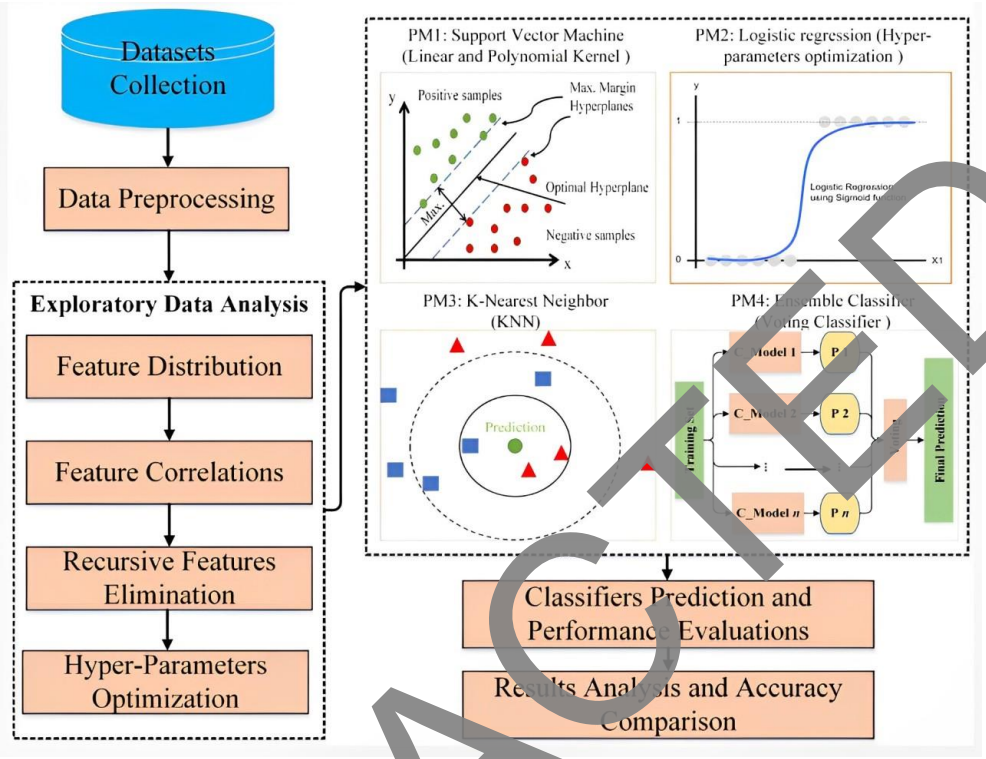

Fig. 4. The above graphic illustrates the systematic process of approach for predicting breast cancer. This strategy involves utilizing data exploratory tools in conjunction with machine learning (ML) classifiers.

3.2 Data Exploratory Techniques (DET)

Data exploration methods, often known as DET, are analytical procedures used to gain insight into the characteristics of a dataset. These methods facilitate the detection of outliers and associated variables that are more readily distinguishable. For researching purpose, the utilized correlation coefficient, feature distribution, and recursive feature removal is important as data exploration strategies.

3.2.1 Correlation of Features

The Pearson Correlation Coefficient "r" [31] computes the association between each pair of the two attributes. Subsequently, the association between two characteristics can be ascertained by classifying them into three categories: features that exhibit positive correlation, features that exhibit negative correlation, and features that exhibit no correlation. If the variables exhibit a positive correlation, the features will exhibit a positive correlation (r = +1). Conversely, if these traits travel in the other direction, they will exhibit a negative correlation {r = −1}.

3.2.2 Recursive features elimination (RFE)

Random Feature Extraction (RFE) is a fundamental procedure in the field of machine learning. Given the dataset's abundance of characteristics, it is crucial to carefully choose the

ideal number of features to enhance the model's performance in making accurate predictions. The essence of doing Recursive Feature Elimination (RFE) is to utilize a reduced number of features that offer enhanced comprehension. The loop will iterate until it can identify the optimal number of characteristics. The characteristics "support_" and "ranking_" were used to indicate the ranking function of the i-th function and to cover the chosen features. Recursive Feature Elimination (RFE) operates by means of iteratively deciding on a subset of functions from the training dataset and regularly removing features until the required range is reached. This is executed via incorporating the specified device getting to know set of rules into the coronary heart of the version, assessing the significance of every characteristic, removing the least full-size ones, and then retraining the model. This process is iterated until a pre-hooked up quantity of characteristics is preserved.

3.2.3 Tuning hyperparameters for optimal performance

A more specific approach to the hyperparameter optimization is a high-quality-tuning of the group of the so-called best parameters, regarded as a system learning method. These parameters serve as values to oversee the learning manner, direction and frequency though they are hired. There are different techniques for hyperparameters tuning including random search, grid search, gradient-based optimization, Bayesian search, Genetic Algorithms and Particle Swarm Algorithms amongst others. Here in this investigation, we used grid search optimization for what we have noticed to be fairly successful in optimization functions. The proposed algorithm employs the brute-pressure method to present capability solutions consisting of the grid of parameter values explained by the parameter. Just like in other grid search processes, the primary concern of grid seek is to achieve the best rankings for pass-validation metrics.

3.2.4 K-Nearest Neighbours (KNN)

This model made use of what is called the K- Nearest Neighbours (KNN) algorithm as a prediction model. As mentioned earlier the KNN algorithm was trained to view the output as a target class. The problem was answered or classified, and this was done through a majority vote from the neighbouring parameter 'K' which was a positive integral value counting between 37 and 38. There are different approaches one can decide to use to compute the distance: Manhattan, Euclidean, Cosine, among others [39].

4. Evaluations of results

This particular portion will assess the results of the suggested forecasting models and data encryption techniques and compare them to previous studies.

4.1 Exploratory Data Analysis (EDA)

In the previous part, we discussed methods for exploring data, and the DE method has produced some important findings. Figure 5 presents the dimensions and categories of both datasets, revealing a significant difference between the WDBC and BCCD datasets. The WDBC dataset comprises benign and malignant classes (a), while the BCCD dataset is divided into absence and presence classes (b). Therefore, a thorough analysis of the WDBC dataset will provide deeper insights. This study specifically focused on examining the means, standard errors, worst values, and correlations to present the dataset.

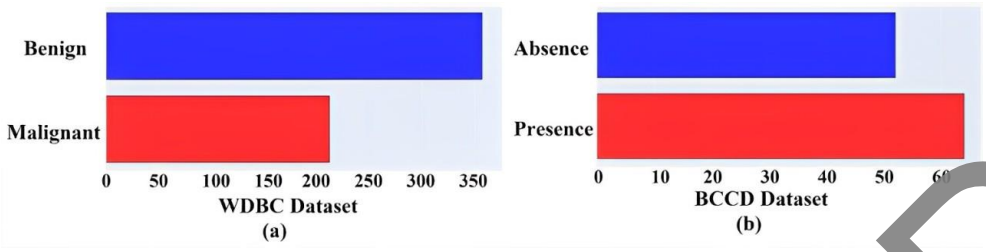

Fig. 5. The breast cancer datasets are classified into two categories: Malignant and Benign for the WDBC classification, Presence and Absence for the BCCD classification. The class distributions of these datasets are shown by the number of samples.

4.2 Evaluations of Predictive Models

The following are the assessments of the provided prediction models (PM):

4.2.1 Prediction Models — SVM

The Linear and Polynomial Kernels, which are used in assisted vector machines, were also covered in this examination. With the use of confusion matrices, Table 1 displays the performance evaluation of each Support Vector Machine (SVM) kernel. The numbers in aspiration indicate the highest performance subliminally. For every training and test set of WDBC data, it was found that the polynomial kernel performed better than the linear kernel. The polynomial kernel fixed the precision rating at about the same level as the linear kernel in the training sets. However, after expanding its audience and developing new offers, it was able to achieve a 98 percent rise in the number of students' grades. 3% F1 score and a 99-time reduction compared to the existing comparable algorithms. 12% accuracy score. In addition, the linear kernel demonstrated satisfactory performance on both the education and checking out datasets. Additionally, the model's overall adaptability of the SVM on the BCCD dataset was found to be less than optimal. Currently, the user-provided text is lacking in textual content. A linear Support Vector Machine (SVM) achieved an accuracy of 76. A polynomial kernel attained only 76% for the F1 measure in this dataset, while a linear kernel obtained an F1 measure of 91%. However, this difference in overall performance is not always statistically significant for the purpose of cancer detection. Therefore, the BCCD dataset was not subjected to any additional assessment, and only the WDBC dataset was used for the final part of the experiment.

Table 1. Comparative analysis of SVM kernels (Linear and Polynomial) on training and testing datasets of WDBC and BCCD to evaluate performance.

Datasets Names	Data Distribution	SVM Kernels	Confusion Matrices			
			P	R	F1 Score	Accuracy
WDBC	Training Dataset	Linear SVM	98.95	98.22	98.57	98.68
		Polynomial SVM	98.62	100	99.3	99.12
	Testing Dataset	Linear SVM	97.14	95.77	96.45	95.61
		Polynomial SVM	97.26	100	98.61	98.25
BCCD	Training Dataset	Linear SVM	72.39	81.05	76.48	76.91
		Polynomial SVM	75.39	75.67	75.53	75.35
	Testing Dataset	Linear SVM	79.01	68.34	73.29	72.09
		Polynomial SVM	74.51	79.29	76.83	76.42

4.2.2 Prediction Models — LR

Three distinct experiment kinds were used in these studies to evaluate the effectiveness of the suggested logistic regression (LR) model. To be more explicit, the three models that were tested were the standard logistic regression (LR model), the logistic regression model that achieved 100% recall accuracy, and the LR-RFE approach. A standard medical dataset, the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, was used for the research. A word of caution: the minimal quantity of training data that is currently available may be the reason why initial training scores were higher than cross-validation scores. To maximize the likelihood of generalizing cross-validation score and training score, it is, however, highly promising to increase the number of training sets. This is due to the LR with RFE yielded the best performance with an F1 & accuracy rate of 0.97. 36% and 97. 04%, respectively. We found that the performance of the standard logistic regression model is slightly inferior than the L 1 logistic regression with a little reduction in the matrix score. The marks belonged to LR were 100% recall; thus, it got to the maximum unfavourable points possible.

Table 2. The efficacy of logistic regression was evaluated using three different approaches: basic logistic regression, logistic regression with 100% recall prediction, and logistic regression with recursive feature elimination (RFE) During the process of cross-validation.

Matrices	Elementary LR	LR Predication with 100% Recall	RFE with LR
Recall	93.87	100	96.22
Precision	97.15	86	98.58
Accuracy	96.66	93.9	98.06
F1 score	95.45	92.5	97.36

4.2.3 Prediction Models — KNN

The KNN prediction version has been examined the use of techniques: primary KNN and KNN with hyperparameters. The basic KNN algorithm functions autonomously with default parameters and presents the outcomes. Conversely, hyperparameters enable the adjustment of parameters for KNN through tuning. The basic KNN model achieved an F1 score of 94.73% and an accuracy of 95.43%. Meanwhile, the K-nearest neighbours (KNN) algorithm with optimized hyperparameters earned an F1 score of 96.35% and an accuracy of 97%.

4.2.4 Prediction Models — EC

Tab-3 shows the results of the evaluation of Ec The outstanding entries serve as representations of the best performances. For the WDBC dataset, three evaluation procedures are employed: ensemble learning, voting classifier (CV), and predicted CV with 100% overlook. Effects stated for the pass-validation (CV) prediction with a recall cost of 80% The collection of logistic regression (LR), pass-validation (CV), and 100% recall cost predicted showed percentage higher impacts. Additionally, it showed that the suggested CV model performed better, with accuracy as high as 96%. 75% accuracy in identifying kidnapping, whereas the corresponding figures for crime detection were 96 and 02% F1 score. supplied and have effectively attained the program's 6% accuracy threshold.

Table 3. Comparative analysis of the performance of Ensemble Logistic Regression, Voting Classifier (Cross-Validation), Voting Classifier Prediction achieved a perfect recall rate of 100%.

Matrices	CV Predication with 100% Recall	Ensemble LR	Voting Classifier
Recall	100	95.75	95.75
Precision	82.7	93.33	96.32
Accuracy	92.1	97.01	**97.61**
F1 score	90.5	95.99	**96.02**

4.3 Comparative analysis of the classifier

The assessment uses cross validation matrices to provide a thorough analysis. Of the three, the one with the lowest F1 score and accuracy (less than 95%) was the CV prediction using LR models and simple KNN, both of which had 100% retrieve. Nevertheless, the comparison showed that, with an accuracy of almost 97% in this study, LR combined with RFE produced superior results. For big data analytics, an accuracy of 04% and an F1 evaluation of 97 are reported. 36 percent. When it comes to choosing the optimal hyperparameters, Polynomial SVM, CV, and KNN all offer advantages over HMM in terms of accuracy. Therefore, as these evaluations demonstrate, every cross-validation compared the LR's performance with RFE to technique had a satisfactory performance, with the greatest accuracy rate of 97.04% during cross-validation. On the other hand, the KNN algorithm with hyperparameter took the longest time to execute, specifically 4.023 seconds, despite achieving a more advanced accuracy of 96.35%.

Table 4. Comparison of the execution duration of each model, along with the highest accuracy attained.

(PM)	Classifiers with Proposed Approaches	Accuracy	Time (s)
PM1: SVM	Linear SVM	98.68	**0.03**
	Polynomial SVM	**99.03**	**0.03**
PM2: LR	Basic LR	96.66	0.266
	LR Predication with 100% Recall	92.9	0.87
	RFE with LR	**98.06**	0.483
PM3: KNN	Basic KNN	95.43	0.031
	KNN Performance with hyperparameter	**97.35**	4.023
PM4: Ensemble Classifier	Ensemble LR	97.01	0.634
	CV Prediction with 100% Recall	92.1	1.845
	Voting Classifier (CV)	**97.61**	0.611

4.4 Comparative analysis with prior research

For the purpose of comparison, it has been compared with prior work which have used the very same WDBC datasets and have offer the most successful results. The comparison is outlined in table 5 where we have provided details of the models/methodologies used in other studies together with the accuracy levels they obtained and the new proposed prediction models together with their outputs. In contrast to the two descriptors, the entries in bold suggest superior outcomes to the studies. An SVM polynomial kernel model was optimized for the dataset and was able to achieve 99% accuracy. 03 % with the total reduction in errors for the LR with RFE model being the closest possible at 97%. 04% [20]. According to a comparative examination, the research's concepts produced new, accurate prediction models that outperformed earlier techniques for accurately diagnosing breast cancer. The new research differs from previous works in that it employs advanced data mining techniques along with appropriate machine learning prediction models, which serves as explanation for these gains. It was suggested that applying DE techniques would enable the maximum level of precision to be attained in the shortest amount of time.

Table 5. Comparing the accuracy regarding breast cancer (BC) Models for forecasting in line with other investigations that employed the identical WDBC dataset.

Author Name	Reference	Year	Model/Method	Best Observed Accuracy
Maglogiannis	[43]	(2007)	SVM Gaussian RBF	97.54%
Mert	[15]	(2015)	KNN	92.56%
Hazra	[29]	(2016)	Utilizing a SV ML with a set of 19 distinct characteristics	94.4%
Osman	[44]	(2017)	SVM	95.2%
Wang	[30]	(2018)	SVM based ensemble learning	96.6%
Abdar	[16]	(2018)	Nested Ensemble 2-MetaClassifier (K)	97%
Mushtaq	[18]	(2019)	KNN with multiple distances (Correlation K=2)	91.1%
Chakravarthy & Rajaguru	[17]	(2019)	KNN Euclidean distance	95.6%
Vijayakumar & Durgalakshmi	[28]	(2019)	SVM	75%
Khan	[45]	(2020)	SVM	97.04%
Shatnawi & Al-Azzam	[34]	(2021)	LR with the calculation of the area under the curve	96%
Predictive Models Accuracy		2022	LR with RFE	97.04%
			Polynomial SVM	99.03%
			KNN Performance with hyperparameter	96.35%
			Voting Classifier (CV)	96.6%

5 Discussion

The majority of our outcome's evaluations focused on analyzing our findings based on the F1 score. We have identified some observations that exhibit considerable disparities in the distribution of features across different classes. For instance, the convexity described in Additional Note 01 exhibited a notable disparity in the allocation between benign and malignant categories. The methods for resampling, namely overestimating, under-sampling, and cross-validations, were employed to achieve balance among these features. The oversampling strategy replicates instances from the minority classes, resulting in an overfitting problem for machine learning algorithms. Conversely, the under-sampling strategy eliminates the majority classes, hence disregarding valuable data. These drawbacks can decrease the accuracy of machine gaining knowledge of in unique packages, consisting of fraud detection, facial popularity, ailment identification, and others. However, the author [48] counseled employing move-validations as the primary approach to tackle the hassle of

imbalanced elegance distribution. Cross-validation use wonderful subsets of the data to evaluate and educate a model. This painting utilized the move-validation technique, specially employing the ok-fold and GridSearchCV techniques, to ensure an equitable illustration of each malignant and benign characteristics in the schooling and trying out dataset. The pass-validations matrices, which blanketed F1 rating, precision, and remember, have been in comparison due to the fact they efficaciously utilized the vital values of TN, FP, TP, and FN to deal with the actual and anticipated training.

By training polynomial SVM, F1 = 0.980 was obtained which indicates high accuracy of the model. 3% and therefore, it was possible to conclude that our division of the symptoms of this type of cancer was correct and the prediction of the tumour was accurate. Therefore, a higher F1 score is indicative of greater diagnostic success rates in malignancies. The F1 score and accuracy of the machine learning algorithm validated in this study are compared to previous studies that employ the same WDBC dataset. Utilizing data mining techniques, these predictive models would aid data analysts in identifying malignant masses by evaluating cancer-related data. Table 4 displays the execution time of each ML model, considering both the minimum and maximum accuracy. This is important because time complexity is a crucial factor for these models. Therefore, based on the aforementioned research, our proposed prediction models and approaches can significantly benefit the cancer field by providing accurate and satisfactory results for diagnosing breast cancer. This work successfully achieved the target of accurately detecting breast cancer using machine learning (ML) models. Unfortunately, we were incapable to determine the exact cause of the malignant features, as this requires the expertise of a specialist in the field. Our prediction models yielded unsatisfactory results with the BCCD dataset, save for SVM. Thus, we omitted those findings from this research. The links/sources of the datasets are available in the "Data Description" column. Due to the datasets being specific to American patients, the findings may not be applicable or efficacious when applied to Asian patients' data. Such a disadvantage relates to the fact that in the future, this particular study may transfer the analysis in a new direction using another dataset, and incorporating a neural network/system.

6 Conclusion

The healthcare profession continues to face significant challenges in accurately and promptly diagnosing various diseases, such as breast cancer, which is crucial for effective treatment. An accurate examination of cancer characteristics remains a laborious and demanding endeavour, mostly because to the abundance of data and the absence of data mining approaches equipped with suitable machine learning classifiers. This work introduced a set of advanced data exploration techniques consisting of four layers. These approaches were applied using four distinct machine learning predictive models: Support Vector Machines (SVM), k- Nearest Neighbours (KNN), Logistic Regression (LR), and an ensemble classifier. The objective was to identify and categorize breast cancer tumours as either benign or harmful. The user's text is empty. The main goal of this study was to apply DE approaches prior to running machine learning ML classifiers on the BCCD and WDBC datasets. By utilizing these mining methodologies, we were able to enhance the performance of the prediction model, achieving an accuracy score and maximum F1 score that surpasses the previous results. The awesome discovery discovered that the preliminary prediction version, making use of an SVM polynomial kernel, had the best stage of accuracy at 98.3%. Furthermore, the utilization of logistic regression with recursive characteristic elimination yielded an accuracy of 97.04%, indicating that the employment of DE approaches correctly identifies more degrees of accuracy. The effects of our have a look at display the effectiveness of our breast most cancers prediction fashions and yield excellent outcomes, completed with a short education length. The utilization of superior models, methodologies, and findings

might permit the medical doctor and data analyst to appoint a greater smart classifier for diagnosing breast most cancers traits. Given the availability of picture data on breast cancer (BC), we will employ deep learning models to identify breast cancer. To address the limited and varied data, we will utilize innovative data augmentation methodologies and data exploration tools. In the future, we will do tests using datasets obtained from other nations to investigate the impact of patients' data from other regions on the performance of the model.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin*. (**2021**), *71*, 209–249.
2. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer statistics, (2022). *CA Cancer J. Clin.* (**2022**), *72*, 7–33.
3. Leão, D.C.M.R.; Pereira, E.R.; Pérez-Marfil, M.N.; Silva, R.M.C.R.A.; Mendonça, A.B.; Rocha, R.C.N.P.; García-Caro, M.P. The Importance of Spirituality for Women Facing Breast Cancer Diagnosis: A Qualitative Study. *Int. J. Environ. Res. Public Health* (**2021**), *18*, 6415.
4. Subashini, T.S.; Ramalingam, V.; Palanivel, S. Breast mass classification based on cytological patterns using RBFNN and SVM. *Expert Syst. Appl.* (**2009**), *36*, 5284–5290.
5. Abdar, M.; Zomorodi-Moghadam, M.; Zhou, X.; Gururajan, R.; Tao, X.; Barua, P.D.; Gururajan, R. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognit. Lett.* (**2020**), *132*, 123–131.
6. Rasool, A.; Tao, R.; Kashif, K.; Khan, W.; Agbedanu, P.; Choudhry, N. Statistic Solution for Machine Learning to Analyze Heart Disease Data. In Proceedings of the (2020) 12th International Conference on Machine Learning and Computing, Shenzhen, China, 15–17 February (2020); pp. 134–139.
7. McWilliam, A.; Faivre-Finn, C.; Kennedy, J.; Kershaw, L.; Van Herk, M.B. Data mining identifies the base of the heart as a dose-sensitive region affecting survival in lung cancer patients. *Int. J. Radiat. Oncol. Biol. Phys.* (**2016**), *96*, S48–S49.
8. Park, K.H.; Batbaatar, E.; Piao, Y.; Theera-Umpon, N.; Ryu, K.H. Deep Learning Feature Extraction Approach for Hematopoietic Cancer Subtype Classification. *Int. J. Environ. Res. Public Health* (**2021**), *18*, 2197.
9. Park, E.Y.; Yi, M.; Kim, H.S.; Kim, H. A Decision Tree Model for Breast Reconstruction of Women with Breast Cancer: A Mixed Method Approach. *Int. J. Environ. Res. Public Health* (**2021**), *18*, 3579.
10. Bicchierai, G.; Di Naro, F.; De Benedetto, D.; Cozzi, D.; Pradella, S.; Miele, V.; Nori, J. A Review of Breast Imaging for Timely Diagnosis of Disease. *Int. J. Environ. Res. Public Health* (**2021**), *18*, 5509.
11. Akay, M.F. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst. Appl.* (**2009**), *36*, 3240–3247.
12. Furey, T.S.; Cristianini, N.; Duffy, N.; Bednarski, D.W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* (**2000**), *16*, 906–914.
13. Zheng, B.; Yoon, S.W.; Lam, S.S. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst. Appl.* (**2014**), *41*, 1476–1482.

14. Seddik, A.F.; Shawky, D.M. Logistic regression model for breast cancer automatic diagnosis. In Proceedings of the (2015) SAI Intelligent Systems Conference (IntelliSys), London, UK, 10–11 November (2015); pp. 150–154.
15. Mert, A.; Kılıç, N; Bilgili, E.; Akan, A. Breast cancer detection with reduced feature set. *Comput. Math. Methods Med.* (**2015**), *2015*, 265138.
16. Abdar, M.; Yen, N.Y.; Hung, J.C.S. Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. *J. Med. Biol. Eng.* (**2018**), *38*, 953–965.
17. Rajaguru, H. Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. *Asian Pac. J. Cancer Prev. APJCP* (**2019**), *20*, 3777.
18. Mushtaq, Z.; Yaqub, A.; Sani, S.; Khalid, A. Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. *J. Chin. Inst. Eng.* (**2020**), *43*, 80–92.
19. Kamyab, M.; Tao, R.; Mohammadi, M.H. Sentiment Analysis on Twitter. In Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality—AIVR (2018), Taichung, Taiwan, 10–12 December (2018).
20. Brause, R.W. Medical analysis and diagnosis by neural networks. In *Proceedings of the International Symposium on Medical Data Analysis, Madrid, Spain, 8–9 October (2001)*; Springer: Berlin/Heidelberg, Germany, (2001); pp. 1–13.
21. Huang, C.L.; Liao, H.C.; Chen, M.C. Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Syst. Appl.* (**2008**), *34*, 578–587.
22. Polat, K.; Günes¸, S. Breast cancer diagnosis using least square support vector machine. *Digit. Signal Process.* (**2007**), *17*, 694–701.
23. Prasad, Y.; Biswas, K.K.; Jain, C.K. SVM classifier based feature selection using GA, ACO and PSO for siRNA design. In *Proceedings of the International Conference in Swarm Intelligence, Beijing, China, 12–15 June (2010)*; Springer: Berlin/Heidelberg, Germany, (2010); pp. 307–314.
24. Muzammal, M.; Qu, Q.; Nasrulin, B. Renovating blockchain with distributed databases: An open source system. *Future Gener. Comput. Syst.* (**2019**), *90*, 105–117.
25. Lim, J.; Sohn, J.; Sohn, J.; Lim, D. Breast cancer classification using optimal support vector machine. *J. Korea Soc. Health Inform. Stat.* (**2013**), *38*, 108–121.
26. Yang, L.; Xu, Z. Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning. *Int. J. Mach. Learn. Cybern.* (**2019**), *10*, 591–601.
27. Fu, Y.; Jung, A.W.; Torne, R.V.; Gonzalez, S.; Vöhringer, H.; Shmatko, A.; Gerstung, M. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* (**2020**), *1*, 800–810.
28. Durgalakshmi, B.; Vijayakumar, V. Feature selection and classification using support vector machine and decision tree. *Comput. Intell.* (**2020**), *36*, 1480–1492.
29. Hazra, A.; Mandal, S.K.; Gupta, A. Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms. *Int. J. Comput. Appl.* (**2016**), *145*, 39–45.
30. Wang, H.; Zheng, B.; Yoon, S.W.; Ko, H.S. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur. J. Oper. Res.* (**2018**), *267*, 687–699.
31. Rasool, A.; Jiang, Q.; Qu, Q.; Kamyab, M.; Huang, M. HSMC: Hybrid Sentiment Method for Correlation to Analyze COVID19 Tweets. In *Advances in Natural*

Computation, Fuzzy Systems and Knowledge Discovery; Springer International Publishing: Berlin/Heidelberg, Germany, (2022); pp. 991–999.

32. Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genom. Proteom.* (**2018)**, *15*, 41–51.
33. Tolles, J.; Meurer, W.J. Logistic regression: Relating patient characteristics to outcomes. *JAMA* (**2016)**, *316*, 533–534.
34. Al-Azzam, N.; Shatnawi, I. Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. *Ann. Med. Surg.* (**2021)**, *62*, 53–64.
35. Khandezamin, Z.; Naderan, M.; Rashti, M.J. Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. *J. Biomed. Inform.* (**2020)**, *111*, 103591.
36. Hasan, A.S.M.T.; Sabah, S.; Haque, R.U.; Daria, A.; Rasool, A.; Jiang, Q. Towards Convergence of IoT and Blockchain for Secure Supply Chain Transaction. *Symmetry* (**2022)**, *14*, 64.
37. Mejdoub, M.; Amar, C.B. Classification improvement of local feature vectors over the KNN algorithm. *Multimed. Tools Appl.* (**2013)**, *64*, 197–218.
38. Yu, Z.; Chen, H.; Liu, J.; You, J.; Leung, H.; Han, G. Hybrid *k*-nearest neighbor classifier. *IEEE Trans. Cybern.* (**2015)**, *46*, 1263–1275.
39. Mondéjar-Guerra, V.; Novo, J.; Rouco, J.; Penedo, M.G.; Ortega, M. Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers. *Biomed. Signal Process. Control* (**2019)**, *47*, 41–48.
40. Pławiak, P. Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals. *Swarm Evol. Comput.* (**2018)**, *39*, 192–208.
41. Bunterngchit, C.; Leepaitoon, S. Simulation-Based Approach for Reducing Goods Loading Time. In Proceedings of the (2019) 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain, 15–17 April (2019).
42. Jafarzadeh, H.; Mahdianpari, M.; Gill, E.; Mohammadimanesh, F.; Homayouni, S. Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation. *Remote Sens.* (**2021)**, *13*, 4405.
43. Maglogiannis, I.; Zafiropoulos, E.; Anagnostopoulos, I. An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. *Appl. Intell.* (**2009)**, *30*, 24–36.
44. Osman, A.H. An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. *Int. J. Adv. Comput. Sci. Appl.* (**2017)**, *8*, 158–165.
45. Khan, F.; Khan, M.A.; Abbas, S.; Athar, A.; Siddiqui, S.Y.; Khan, A.H.; Hussain, M. Cloud-based breast cancer prediction empowered with soft computing approaches. *J. Healthc. Eng.* (**2020)**, *2020*, 8017496.
46. Google Search" [https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning], [https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm]
47. Sandip S. Panesar, Rhett N. D'Souza, Fang-Cheng Yeh. Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in a Small, Heterogeneous Glioma Database Aprril (2019), 10001 [https://doi.org/10.1016/j.wnsx.2019.100012]

48. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, (2013).