Breast Cancer Detection Analysis Using Different Machine Learning Techniques: South Iraq Case Study

Salma Abdulbaki Mahmood[1], Myssar Jabbar Hammood Al-Battbootti[2], Saad Shaheen Hamadi[3], Iuliana Marin[4], Costin-Anton Boiangiu[2], and Nicolae Goga[4]

[1]*Department Computer Information Systems, Computer Science and Information Technology, University of Basrah, Basrah, 61004, Iraq*

salma.mahmood@uobasrah.edu.iq

[2]*Faculty of Automatics and Computer Science, National University of Science and Technology Politehnica Bucharest, Splaiul Independentei 313, 060032 Bucharest, Romania*

myssar_jabbar.al@stud.fils.upb.ro
costin.boiangiu@cs.pub.ro

[3]*Faculty College of Medicine, University of Basrah, Basrah, 61004, Iraq*

saad.shaheen@uobasrah.edu.iq

[4]*Faculty of Engineering in Foreign Languages, National University of Science and Technology Politehnica Bucharest, Splaiul Independentei 313, 060032 Bucharest, Romania*

iuliana.marin@upb.ro
n.goga@rug.nl

ABSTRACT

Contemporary oncology has seen a growing interest in digital technologies, whose integration with extensive healthcare and clinical data has raised new aspirations in managing patient profiles and organizing treatment plans. Among the commonly used digital technologies are Machine Learning (ML) methods that can perform many tasks, such as prediction, classification, and description, based on previously stored big data with high precision and speed. This study aims to develop a predictive ML model for early prediction of breast cancer based on a set of medically categorized risk factors. The locally collected database contained 415 instances from Al-Sadr Teaching Hospital in Basrah, Iraq, 219 (53%) of which were breast cancer patients, whereas 196 (47%) of them were control, respectively non-patients. It trained seven machine learning methods, namely Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logical Regression (LR), Multinominal Naïve Bayes (NB), and Gaussian NB. The dataset was cleaned and balanced before being used. The results proved the superiority of the Decision Tree model with 96% accuracy, 96% sensitivity, and 96% specificity, the Random Forest model with 94% accuracy, 100% sensitivity, and 87% specificity, and SVM model with 92% accuracy, 96% sensitivity, and 87% specificity, respectively. Other models gave diverging results. The current study concluded that modern technologies should be employed to raise awareness and control diseases. The need to adopt Electronic Health Records (EHR) to ensure the integration of clinical data of different types recorded over time for patients contributes to building accurate and reliable prediction models.

1. INTRODUCTION

Breast cancer is the most common type of cancer among women worldwide, especially in the Middle East and North Africa, constituting about 53%, which leads to death, with an estimated death rate of 245,000 women in 2015, as based on the research of Azamjah, Soltan-Zadeh, and Zayeri (2019). In this regard, the number of newly diagnosed cancer cases will surpass an incidence rate of 98.41 per 100,000 individuals by 2030, according to Zhang et al. (2024). Only in Iraq, from the year 2000 to 2019, the incidence rate of new cases of cancer has been trending upward, from 52.00/100,000 cases to 91.66/100,000 cases, respectively based on the studies of Mohsin and Mohamad (2024). Breast cancer in Iraq remains the leading cause of cancer-related death in women. It constituted about one-

https://doi.org/10.36001/IJPHM.2025.v16i1.4240

third of all registered cancer cases within the country in 2019, according to Al-Hashimi (2021). Based on the research, the highest percentage and incidence rate of the top ten cancers in 2019 was breast cancer, 34.08% and 35.95/100,000, respectively. Mammography is the most widely used method for detecting early-stage breast cancer, and it is the recommended standard for periodic breast cancer screening. The detection of breast cancer using machine-learning-based techniques has gained significant attention in recent times.

Machine-learning models have long been used in CAD as classifiers, regression models, feature mapping algorithms, data enhancement, and image segmentation to identify subtle objects in complex backgrounds that a human reader may miss or misinterpret, based on the review of Syamsiah Mashohor et al. (2023). A set of machine learning algorithms and feature selection methods has been established by Dar, Rasool, and Assad (2022). The machine learning methods developed have shown acceptable performance on the task, with training and testing on the same database. This study aims to create a robust and reliable ML-based methodology for breast cancer detection by conducting a comprehensive evaluation and validation of different algorithms. A comparative assessment is needed to identify the optimal approach. In the current article, we evaluated various ML algorithms: Decision Tree, Random Forest, Support Vector Machines, K-Nearest Neighbors, Logical Regression, Multinominal Naïve Bayes, and Gaussian Naïve Bayes.

Whereas most of the existing literature in this area heavily relies on popular, well-established datasets such as Wisconsin Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC), as found by Yadav, Singh, and Kashtriya (2023), our study intentionally steps away from this common practice. We decided to employ a local dataset for training our algorithms and, in doing so, have chosen a methodology with numerous ensuing benefits. Utilizing a locally collected dataset from Al-Sadr Teaching Hospital in Iraq, this study pursues the following objectives:

- To determine the optimal algorithms and perform a comparative analysis of the classification performance of leading ML models for breast cancer diagnosis;

- In contextual relevance, local datasets provide information that is, by definition, specific to the population or health care system under study. This contextual congruence allows the findings gleaned from the data to be generalized directly to the population of interest and hence enhances the ecological validity of the study;

- Data quality and completeness, by diverging from the commonly used datasets, our study contributes to the diversification of the knowledge base in this field.

This research concludes that over the latest years, the incidence of breast cancer has risen in Iraq and thus is an important health issue. There is an urgent need for action at the national level to be taken so that better assessment and management can be done.

2. RELATED WORK

Breast cancer studies differ in using diverse datasets but are similar in employing the most common and widely used machine learning methods, according to Jain & Singla (2023). Thus, some of them used only demographic risk factors such as lifestyle and laboratory data for training ML to predict breast cancer. Many developed ML models based on mammographic stereotypes or data biopsy. At the same time, others used genetic data to predict breast cancer. In recent years, there has been a variety of applications that depend on a different dataset.

Mohaimenul & Poly (2019) focused on the construction of a breast cancer risk prediction model. They used the most used machine learning models such as Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Artificial Neural Network (ANN), and 10-Fold Cross-Validation. The dataset comprised 116 patient records with ten clinical features (age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resisting, and MCP-1). They proved that the KNN method outperformed the other, using the area under the receiver operating curve, sensitivity, and specificity of 0.95, 0.80, and 0.91. respectively. They determined the limitations of their system in points using deterministic features, local datasets, and small sample sizes.

Ferroni et al. work (2018) assessed the risk of disease progression in an oncology setting of breast cancer patients. They used a dataset of 454 samples distributed as a training set (n=318). A testing set (n=136), integrating multiple clinicopathological features and genomic data, was stored in patients' EHR and performed Bayesian analysis method with positive (+LR) and negative (−LR) likelihood ratios were used to estimate the probability of breast cancer progression and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) to evaluate their model. They used combined ML (SVM, Neural Networks, Naïve Bayes) and receiver operating techniques and a 3-fold cross-validation technique to investigate whether the combined decision support system (DSS) could distinguish between recurrent and non-recurrent patients. The study confirmed that there are no models that significantly outperformed others. The recorded limitations are mono-institutional, and the sample size of the used dataset is too small.

Battineni, Chintalapudi, and Amenta (2020) developed a tumor classification system using limited features of the Wisconsin Breast Cancer Dataset (WBCD), which was obtained from the University of California Irvine (UCI) machine-learning repository. The used dataset involved digitized images of malignant tumors for 569 females,

including 212 (37%) diagnosed as the malignant type and 357 (63%) diagnosed as benign type with 30 cell features.

Three supervised ML algorithms, namely SVM, LR, and KNN, were implemented with statistical analysis to identify highly associated features in malignant classification and 10-fold cross-validation. The researchers did two different experiments to determine total features and limited features. It was showed that both SVM and LR models generated 97.66% accuracy with total feature evaluation, and the SVM accuracy was improved by 98.25% with selective features. The limitations of the study are the small datasets of biomarkers and the necessity to use multi-centric databases.

The study of Naji et al. (2021) aimed to develop, predict, and diagnose breast cancer. They used the Wisconsin Diagnostic dataset to five ML algorithms, namely SVM, RF, LR, DT, KNN, with a feature selection algorithm to minimize features number. It was observed that SVM outperformed all other classifiers and achieved the highest accuracy (97.2%), precision (97.5%), and AUC (96.6%). The WBCD database is a limitation of this work.

Afrash et al. (2022) developed a risk prediction and early warning model of breast cancer. Incorporating genetic algorithm (GA) with ML algorithms, such as KNN, radial basis function (RBF), DT, artificial neural network (ANN), feedforward neural network (FNN), probabilistic neural network (PNN), and pattern recognition were used. Based on the dataset of 3930 cases, 1270 (32.31%) patients and 2660 (67.69%) were diagnosed as non-breast cancer.

With 35 features, including demographical, clinical, and lifestyle variables, the dataset got minimized to ten most important features, namely age, consumption of dairy products, breast cancer family history, breast biopsy, chest X-ray, hormone therapy, alcohol consumption, being overweight, having children, and education statuses, using GA. To balance their dataset and avoid bias problems, they used the Synthetic Minority Over Sampling Technique (SMOTE) and 10-fold cross-validation to avoid overfitting problems. The DT algorithm with 10-fold cross-validation presented the best accuracy of 99.2%, a specificity of 99.5%, and a sensitivity of 97.9%. The limitation is that a single-center dataset with low quality and low quantity was used.

Rabiei et al. (2022) aimed to predict breast cancer using the locally collected dataset in Motamed Cancer Institute, Academic Center for Education, Culture and Research (ACECR), Tehran, Iran. Their dataset included 5,178 records, 25% of which were diagnosed as breast cancer patients, with 24 attributes distributed (eleven demographic features, nine laboratory features, and four mammography features). Various ML methods, such as RF, multilayer perceptron (MLP) neural network, gradient boosting trees (GBT), and GA, were used. SMOTE was used to balance the dataset, 3-fold cross-validation, and genetic algorithm to enhance the performance. The RF method outperformed the other methods with an accuracy of 80%, sensitivity of 95%, specificity of 80%, and an area under the curve of 0.56. The need for extensive datasets from different institutions, for a multi-center study, is considered a limitation of this paper.

Chtouki et al. (2022) aimed to develop a 5-year breast cancer survival prediction system, and binary classification (survival or not survival) was a result. They used the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset, which involved 1904 records comprising 175 gene mutations, 31 clinical characteristics, and the mRNA z-scores for 331 genes. The factors consisted of features, such as prognostic factors, including age, race, marital status, primary site, laterality, behavioral code, histology, tumor size, lymph node, extension, surgery, radiation, and tumor, node, metastasis (TNM) stage. Seven ML models, namely LR, SVM, DT, RD, Extremely Randomized Trees (ERT), and KNN, were used in this study. The work incorporated the SelectKBest method to calculate the correlation between the variables and survival time with the 10-fold cross-validation. Adaptive Boosting (AdaBoost) achieved the highest accuracy, 75.4%. As a drawback, the limited dataset needs to enlarge data samples or use another dataset.

González-Castro et al. (2023) developed a 5-year breast cancer recurrence prediction system. The dataset was collected in Centre Hospitalier Universitaire (CHU) de Liège, Belgium, which contained 823 instances having both structured and unstructured data from EHR. SMOTE was applied to balanced data. LR, DT, GBT, eXtreme Gradient Boosting (XGB), Deep Neural Network (DNN), and grid-search algorithm are involved in hyperparameters optimization. The XGB model recorded the best model performance with precision of 0.900, recall of 0.907, F1-score of 0.897, and the area under the receiver operating characteristic, AUC ROC of 0.807. The fact of using specific datasets in one medical center is the limitation of the study.

Chen et al. (2023) presented a system for the early diagnosis of breast cancer depending on WBCD, which was obtained from the UCI machine-learning repository. This dataset involved digitized images of malignant tumors. It contained 569 female instances, including 212 (37%) diagnosed as malignant type and 357 (63%) as benign type and 30 features. ML algorithms such as XGBoost, RF, LR, and KNN were used, and other methods such as standardization method to eliminate the impact of different dimensionality, the Pearson correlation test to reduce features to 15 features, and a stratified sampling method to solve the problem of classes imbalance.

XGBoost model outperformed recall, precision, accuracy, and F1-score at 1.00, 0.960, 0.974, and 0.980, respectively. Using structured and unstructured data causes a high cost of

data preprocessing and training, so an NLP-based approach is needed.

Poornajaf and Yousefi (2023) used the WBCD dataset from the UCI repository, which included 699 samples, to build a system for breast cancer prediction. Several ML algorithms were exploited to diagnose the cancer tumor, such as LR, DT, RF, and KNN, with feature selection and cross-validation methods to improve the results. The best model was the LR algorithm with an accuracy value equal to 99.14% and AUC ROC equal to 99.6%.

Parekh and Dahiya (2023) presented a system for correctly early prediction of breast cancer. A dataset called Curated Breast Imaging Subset (CBIS) of digital screening mammography was used. CBIS comprises 2620 scanned film mammography, classified as "B" and "M" breast cancers, for exactly 1319 patients. SVM, LR, KNN, Naïve Bayes Classifier, DT Classifier, and two ensemble algorithms, AdaBoost and XGB, were used. The XGB classifier was the best model with an F1-score of 0.8912, which is better than every other algorithm. The recorded limitation is that it was not multimodal.

Iparraguirre-Villanueva et al. (2023) developed a model for diagnosing or predicting the probability of breast cancer in patients. The Wisconsin repository, which includes 569 observations and 32 features, was used. MLP, KNN, AdaBoost, Bagging, GBT, and RF were ML methods used in their study, with correlation and ensemble methods to enhance the resulting model's performance. The RF, GB, and AdaBoost models achieved 100% accuracy, outperforming the others. The limitations of this work include, as the researcher noted, imbalance classes which were found, limited data, and the need for powerful resources.

Chakkouch et al. (2023) presented a comparison study to find the best model from different machine learning techniques for breast cancer recurrence type prediction. A local clinical and pathological data dataset was collected from a single center in Meknes, Morocco, between 2015 and 2022. The dataset encompassed 1,189 patients who underwent surgery, radiation therapy, and chemotherapy.

The follow-up of at least 60 months, with 19 features, including tumor size, age, hormone receptor status, histological grade, lymph node status, human epidermal growth factor receptor 2 (HER2) status, progesterone receptor (PR) status, estrogen receptor (ER) status, chemotherapy, targeted therapies, radiation therapy, hormonotherapy, healthy eating, physical activity, type of psychosocial stress, and type of recurrence.

LR, DT, KNN, and ANN algorithms were used for building a multi-classification model, including local, regional, mixture of local, regional, and distant recurrence. The comparison yields that the ANN algorithm outperformed the other algorithms with 91% accuracy, followed by the DT algorithm and KNN, which also performed well with accuracies of 90.10% and 88.20%, respectively.

The LR algorithm had the lowest accuracy of 84.60%. The downside of the study is that it focused only on breast cancer recurrence type predication and did not consider other factors such as disease-free survival or overall survival.

3. DATA ACQUISITION AND PREPROCESSING

The dataset utilized in this study, sourced from the Oncology Center at Al-Sadr Teaching Hospital in Basra, initially comprised 768 instances, including 572 malignant and 196 benign instances. Class imbalance, a common occurrence in medical datasets, can significantly affect the performance of machine learning algorithms, particularly when the minority class is underrepresented.

In our study, we encountered a notable class imbalance, with 74.48% of the instances belonging to the malignant class and only 25.52% representing benign cases. This skewed distribution can lead to biased models that favor the majority class, compromising the accuracy of their predictions for the minority class. To address this challenge, we employed the Synthetic Minority Over-sampling Technique (SMOTE) of Adi Pratama and Oktora (2023). However, extensive preprocessing steps were meticulously executed to prepare the dataset for model development. These steps included handling missing values through median imputation, feature normalization to ensure consistent scales, clipping outliers exceeding 3 standard deviations to mitigate their influence, and one-hot encoding of categorical variables to facilitate their incorporation into the models.

In the second stage, following the implementation of preprocessing steps, the refined dataset comprised 415 instances. Of these, 219 instances (53%) represented confirmed breast cancer cases, while 196 instances (47%) served as control cases. This near-balanced distribution between positive and negative cases enhances the reliability of our subsequent analyses and model predictions. Each instance includes eight features represented as risk factors according to medical perspective and studies (Roheel et al., 2023; Cuthrell and Tzenios, 2023; Daly et al., 2021], including breast cancer status (ST), age, body mass index (BMI), age at puberty, menopausal status, marital status, number of children, age at first birth, and family history of breast cancer.

The diversity of the dataset was further enriched by the inclusion of patients spanning a wide age range, from 11 to 96 years. Deliberate sampling measures were undertaken to ensure a balanced representation of cancer-positive and cancer-negative cases. The meanings and value types are described in Table 1.

Feature	Explanation	Value
ST (Status of the individual)	ST indicates whether they have breast cancer (patient) or not (control). This is the target variable in the breast cancer prediction model.	1 for breast cancer, 0 for control
Age	Age is a significant risk factor that increases as a woman gets older. Most cases are diagnosed in women aged 50 and over.	Range from 18-92 years
BMI (Body Mass Index)	BMI is a measure of body fat based on height and weight. Higher BMI, post-menopause, is associated with an increased risk of breast cancer. Obesity influences hormone levels and cancer development.	Range from 13-63
Puberty	The age of menarche influences breast cancer risk. Early puberty leads to longer exposure to estrogen that increase the risk.	Range from 9-19 years
Marital Status	Marital status itself is not a direct risk factor, but it can be associated with factors like socio-economic status, stress levels, and support systems, which can indirectly influence breast cancer risk and health outcomes.	0 for single, 1 for married, 2 for divorced, 3 for widow
Children Number (Children No)	The number of children and the age at which the mother has her first child influence breast cancer risk. Those who have their first child at a younger age have a lower risk of breast cancer due to pregnancy hormonal changes.	Range from 0-15 children
First Birth Age	The age at which a woman has her first child is an important factor. Having a first child at an older age or not having children at all is associated with an increased risk of breast cancer. Early childbirth is protective because it reduces the number of menstrual cycles a woman has in her lifetime.	Range from 0-45 years
Family History	Family history of breast cancer increases the risk. If a close relative has been diagnosed with breast cancer, the risk is higher due to potential inherited genetic mutations (BRCA1, BRCA2).	Range from 0-5 patients.

Table 1. Risk Factors used as Features (Dependent Variables) of ML Algorithms

Table 2 showcases participant distribution based on family history.

Cases per family	Malignant cases	Benign cases	Total
0	145	115	258
1	207	38	245
2	100	22	122
3+	120	22	142
Total	572	196	768

Table 2. Participant distribution based on family history

The distribution of individuals categorized by their family history of cancer is explained in Table 2. Numerous interesting relationships and patterns may be seen in the data. Firstly, it is observed that individuals without a family history of illness accounted for approximately 25% of all cancer cases. which implies that although genetic predisposition plays a major part, other variables like lifestyle decisions, environmental exposures, or spontaneous genetic alterations may also play a significant role in the development of cancer.

In the cohort with no family history, the malignant and benign cases are equally distributed, although there is a slight predominance of malignant cases: 56% as opposed to 44% benign cases. A nearly equal distribution in the absence of familial risk factors calls for a deeper investigation of the underlying mechanisms driving malignancy in these cases. Of particular interest is the pronounced correlation between the number of family members previously diagnosed with cancer and the increased likelihood of an individual presenting with the disease. The data reveal a marked shift in the malignant-benign ratio as the number of affected family members increases:

- In one-member previously diagnosed families, the malignant-to-benign ratio was 85:15.

- Where two members of the family had already been diagnosed, the ratio is slightly moderated but still substantial at 82:18.

- Where three or more family members have been diagnosed, the ratio stays as high as 85:15 - identical to that shown for families with only one affected member.

The input dataset represents a very consistent and significant increase in the percentage of malignant cases associated with a familial history and thus constitutes very strong support for the genetic factor in cancer risk. The most intriguing observation is that the ratios for families with one and with three or more affected members are identical, 85:15 and may suggest a plateau effect beyond a certain threshold in the genetic risk.

	ST	age	BMI	puberty	menopause	marital status	no of children	First birth age	family history
ST	1.00								
age	0.62	1.00							
BMI	0.04	0.08	1.00						
puberty	-0.07	0.03	0.02	1.00					
menopause	0.25	0.59	0.02	0.00	1.00				
marital status	0.46	0.40	-0.02	-0.01	0.12	1.00			
children no	0.48	0.37	-0.25	-0.01	0.13	0.44	1.00		
first birth age	0.36	0.34	-0.05	0.00	0.08	0.45	0.51	1.00	
family history	0.20	0.15	0.07	-0.04	0.07	0.11	0.08	0.09	1.00

Figure 1. Correlation matrix risk factors associated with breast cancer

Figure 1 represents the correlation matrix that employs a color-coding scheme to highlight the significance and directionality of the relationships between the features and the disease status. The diagonal elements, depicted in red, represent the correlation of each feature with itself, which is inherently 1 and therefore not informative for the analysis.

Features exhibiting a correlation coefficient greater than 0.5 with the disease status variable (ST=1 for patients, ST=0 for non-patients) are shown in blue. This positive correlation indicates that as the values of these features increase, the likelihood of breast cancer incidence also increases. These characteristics can be considered highly influential and potentially valuable predictors in breast cancer risk assessment models. Conversely, features with correlation coefficients less than -0.5 are represented in pink, suggesting an inverse relationship with the disease status. As the values of these features decrease, the probability of breast cancer incidence increases. Understanding these inverse relationships can provide insights into potential protective factors or risk-mitigating variables.

Features depicted in green are those with correlation coefficients between -0.5 and 0.5, indicating a weaker or less substantial association with breast cancer incidence. While these features may still contribute to the overall risk assessment, their individual impact on disease occurrence is likely less pronounced compared to the highly correlated features.

By employing this color-coded correlation matrix, the figure aims to provide a concise and visually interpretable representation of the complex relationships between various features and breast cancer incidence, facilitating the identification of key predictive variables and potential areas for further investigation or model development.

4. METHODOLOGY

4.1. Breast Cancer Prediction Methodology

The proposed methodology for breast cancer prediction employs a multi-faceted approach, leveraging advanced machine learning techniques and a comprehensive set of risk factors. Initially, the preprocessed data is partitioned into training (80%) and testing (20%) subsets, following standard practice for model evaluation based on Zhu et al. (2023). Feature selection is performed to identify the most relevant attributes, enhancing model performance and reducing computational complexity (Mueller et al., 2023), utilizing recursive feature elimination (RFE) in conjunction with correlation analysis to identify the most salient predictors (Reshan et al., 2023). This approach aligns with recent studies emphasizing the importance of dimensionality reduction in improving model performance and interpretability (Lee, 2023).

The core of the predictive framework comprises an ensemble of machine learning algorithms, including Decision Trees, Random Forests, and Support Vector Machines, each chosen for their demonstrated efficacy in medical diagnostic applications (Asif et al., 2024), as shown in Figure 2.

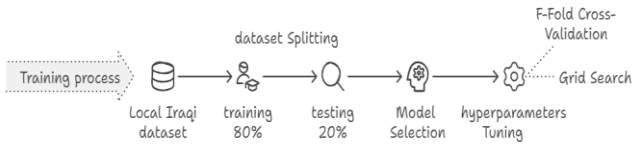

Figure 2. Model training and data processing

Notably, the Decision Tree model exhibited superior performance, achieving an accuracy of 96.1%, which surpasses the benchmark set by comparable studies in the field (Obuchowicz et al., 2024). Furthermore, the integration of k-fold cross-validation ensures robust model evaluation and mitigates overfitting, a critical consideration in developing clinically applicable predictive tools (Darwich and Bayoumi, 2022).

The proposed method of Edwards et al. (2023) incorporates a unique feature engineering approach, deriving complex interactions between risk factors based on domain expertise and recent epidemiological findings. This holistic methodology not only achieves high predictive accuracy but also provides interpretable results, facilitating its potential integration into clinical decision support systems (Garcia-Moreno et al., 2024).

4.2. Feature Importance

This study employed Pearson correlation coefficients to assess the significance of various features in predicting breast cancer risk. Age emerged as the most influential factor (r = 0.72), followed by BMI (r = 0.41), number of children (r = 0.4), age at first birth (r = 0.37), marital status (r = 0.31), and family history (r = 0.23), as illustrated in Figure 3.

Interestingly, puberty factors showed a negative correlation (r = -0.18). These findings largely align with existing literature, confirming age as a critical risk factor and highlighting the importance of reproductive history. The unexpected negative correlation with puberty factors warrants further investigation.

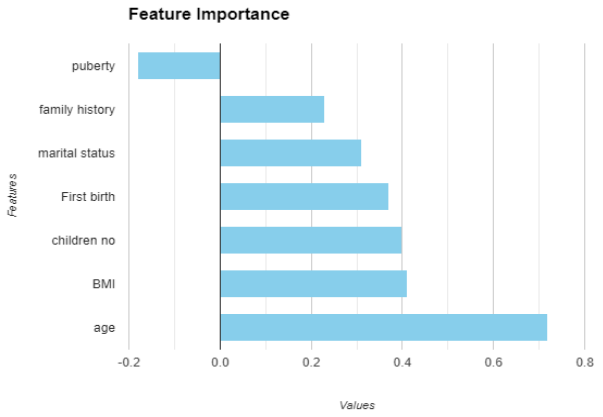

Figure 3. Pearson correlation coefficient between the used features and target class

Recursive feature elimination was used to identify the optimal subset of predictors, balancing predictive accuracy with model simplicity. This analysis provides valuable insights into breast cancer risk factors, corroborating previous research while also revealing areas for future study, particularly regarding the complex interactions between different risk factors across diverse populations.

5. RESULTS AND DISCUSSION

To determine which machine learning algorithm is best for predicting breast cancer, a variety of models were applied to the Breast Cancer Iraqi dataset. The models were then assessed using important performance metrics like the Confusion Matrix, Accuracy, Precision, Sensitivity, F1 Score, and AUC to identify the accurate algorithm for breast cancer prediction.

One of the most used tools to assess classification problems in which the output can belong to two or more classes is the confusion matrix. It is structured in two dimensions, "Actual" and "Predicted", where the matrix entries could be categorized as True Positives, True Negatives, False Positives, and False Negatives. As shown in Table 3.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 3. Confusion matrix components

The most common performance metric for classification models is accuracy. This is given by the ratio of the correct predictions to the total number of predictions. Precision, mainly used for document retrieval and other binary classification tasks, denotes the number of correctly predicted positive instances out of the total instances predicted as positive by the model.

Sensitivity, referred to as recall or the true positive rate, means the ratio of actual positive cases the model correctly identifies. The F1 Score delivers a balanced measure of Precision and Sensitivity, providing their harmonic mean. This is particularly useful when one is dealing with an uneven class distribution, giving a balance to the tradeoff between precision and recall.

Table 4 and Figure 4 present the accuracy rates of various classifiers regarding the Al-Sadr Teaching Hospital Iraqi Diagnostic Dataset. To check for overfitting in our models, training and testing accuracies for all used methods were measured. Among all, the most consistent performance was given by the Decision Tree, which topped with the highest accuracy on the test set with a rate of 96.4%. and 96.1%. for the training set, as a matter of fact, it pinpoints the Decision Tree as the most accurate algorithm in this dataset, thus proving to be very effective in the prediction of breast cancer.

Predictor	Training Accuracy	Testing Accuracy
Decision Tree	0.961	0.964
Random Forest	0.941	0.941
SVM	0.921	0.922
KNN	0.892	0.892
Logical Regression	0.852	0.853
Multinominal NB	0.754	0.755
Gaussian NB	0.751	0.745

Table 4. Training and testing accuracy for all used predictor

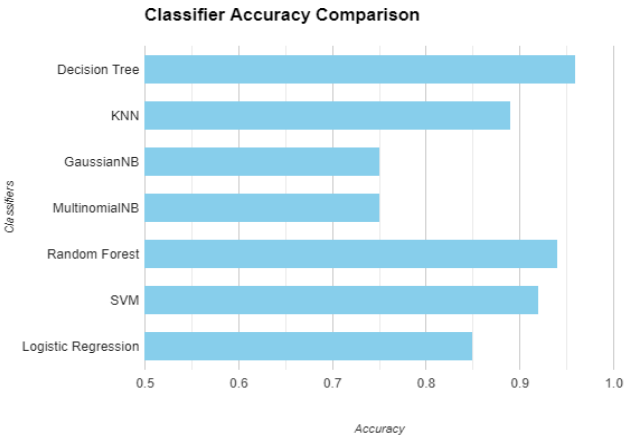

Figure 4. Accuracy metrices for all used ML methods

In addition to sensitivity and specificity, A thorough evaluation of the machine learning model's overall performance is given via the AUC evaluation metric.

A curve line that approaches 1 indicates better model performance, while a value closer to 0 suggests no predictive ability. The decision tree, random forest, and logistic regression classifiers performed the best when compared to other techniques, as shown in Figure 5.

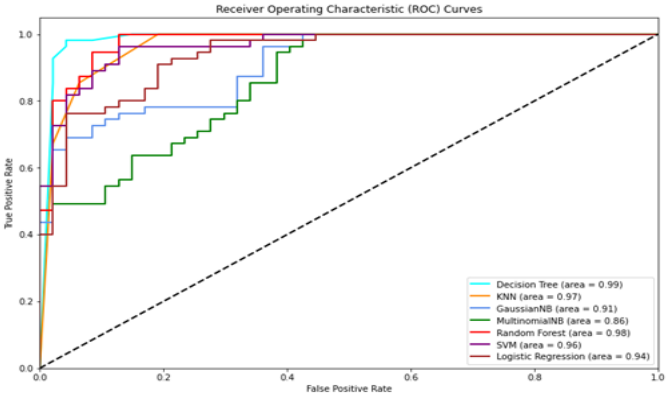

Figure 5. Area under the ROC curve for all used predictors

Other metrics are given in more detail in Tables 5 and 6, where the Decision Tree predictor again outperformed all the rest of the ML methods for being of higher reliability when all metrics evaluated, especially sensitivity and specificity, are concerned.

The fact that the Decision Tree had more balanced values for sensitivity and specificity means it was not only sensitive, noticing the true positives, but also specific, correctly identifying the true negatives. This makes it a very robust choice for this dataset when it comes to predicting the risk of breast cancer. It was followed by Random Forest, which, for the same dataset, is more reliable regarding sensitivity but not as much in terms of specificity. This suggests that Random Forest, while very good at capturing all the true positives, was more prone to false positives than the Decision Tree.

Classifier	Confusion Matrix
Decision Tree	[[45, 2] [2, 53]]
Random Forest	[[41, 6] [0, 55]]
SVM	[[41, 6] [2, 53]]
KNN	[[44, 3] [8, 47]]
Logical Regression	[[37, 10] [5, 50]]
Multinominal NB	[[22, 25] [0, 55]]
Gaussian NB	[[33, 14] [12,43]]

Table 5. Confusion matrix

Classifier	Accuracy	Precision	F1-score	Recall	Specificity
Decision Tree	0.96	0.96	0.96	0.96	0.96
Random Forest	0.94	1.00	0.94	0.87	1.00
SVM	0.92	0.95	0.91	0.87	0.96
KNN	0.89	0.85	0.89	0.94	0.84
Logical Regression	0.85	0.88	0.83	0.79	0.91
Multinominal NB	0.75	1.00	0.64	0.47	1.00
Gaussian NB	0.75	0.73	0.72	0.71	0.78

Table 6. Evaluation metrics for all used classifiers

While our research contained features such as age, BMI, and family history, the sociocultural factors that have been included in this project, like marital status and number of children, have proven to be particularly significant in Middle Eastern countries and generally ignored by Western-centric studies. This dataset will, therefore, become decidedly important for research that deals with similar settings in demographic and cultural aspects.

Our study reported the highest accuracy of 0.96 for the Decision Tree classifier, beating other models like Random Forest at 0.94 and SVM at 0.92. These findings comply with global studies, including that done in the research of Manikandan, Durga and Ponnuraja (2023), using the SEER dataset, reporting high performance of Decision Trees in the outcome prediction of breast cancer. Indeed, Decision Trees can handle complex interactions between the risk factors, hence their high performance with our data set. However, Random Forest models are preferred since they are more robust and generalize much better to other datasets, as revealed in a 2021 meta-analysis of breast cancer prediction models belonging to the research of Li et al. (2021). The study conducted on 550 patients to predict the survival and metastasis of breast cancer had the highest accuracy, standing at 93%, and it belonged to the SVM, which is very close to ours, standing at 92%, as in the work of Tapak et al. (2019).

Similar performances have been reported at regional levels. In 2024, a study in Iran of Dianati-Nasab et al., reported the accuracy of the Random Forests at about 83% on the local dataset, though depending on the features used, it would vary in exact performance. That our accuracy rates are higher can thus be explained based on the peculiar characteristics of our dataset and features not that common in other regional datasets, which are culturally relevant.

The Decision Tree model thus allows both high sensitivity and specificity in the classifying of cases and controls, resting on quite symmetric performance-0.96 each in our study-which is quite important clinically as any misclassification will have serious consequences.

Comparatively, other similar model studies, such as the study on the prediction of breast cancer in Saudi Arabia pf Al-Rikabi and Husain (2012), have reported lower specificities, thereby suggesting additional relevant features in our dataset have contributed to better model performance.

While the imaging techniques in Alotaibi et al. (2024) are mostly dependent on the expertise of the radiologists and quality of imaging, which may be variable across Saudi Arabia, our machine learning models are pre-trained on all with minimal overfitting to balance the performance for both training and testing datasets. This robustness insinuates that the performance of our models will remain high across various datasets, where the accuracy of the imaging could be highly variable depending on equipment and expertise of the radiologist.

With the high performance outlined, especially for our Decision Tree and Random Forest models, clinically, such models can be applied in Iraq and regions of similar environmental settings. Especially, the models containing easily observable risk factors such as marital status and number of children can be of great use in early detection programs. However, generalization of these models to other populations is of prime importance, since their acceptance is limited beyond Iraqi boundaries due to genetic and cultural variations.

Overall, the closeness of the training and testing accuracies throughout all the models speaks to the strength of the modeling process employed in this study. The recommendation of the Decision Tree and Random Forest models is due to their low risk of overfitting with high accuracy. The contribution of this research study lies in the emerging literature body on breast cancer prediction in the Middle East and will be useful in both regional and global studies.

6. CONCLUSION

This study focuses on using machine learning to develop a model capable of predicting the possibility of cancer before its occurrence or early prediction of the disease before its progression by recording simple data that are medically considered to be risk factors for the disease or its development.

This study used locally collected data to obtain seven prediction models, including 408 cases described by eight risk factors and seven machine learning methods. The prediction models were compared using standard evaluation tools to find the best model.

The Decision Tree model proved to be the best model with the most accurate (accuracy 0.961) and stability regarding medical metrics, sensitivity, and specificity. They were followed by the Random Forest model with accuracy (0.941) and the SVM model with the lowest rating, accuracy (0.921). In contrast, the rest of the models fluctuated with different assessment values.

This study is of great importance in spreading health awareness of the risk factors that may cause the incidence of breast cancer to find personalized plans to avoid it on the one hand, and on the other hand, early detection and its significant role in controlling the disease and preparing a health treatment plan early.

One of the challenges of this study is the limited dataset recorded in one hospital for a short period of time, not exceeding two months. It can be suggested to develop the study by increasing the recorded risk factors from several areas in Basra governorate and in other governorates of Iraq to include environmental factors, living habits, and nutrition among these risk factors. The study also proposes the adoption of a patient health record that collects all the information required for a reliable prediction process.

ACKNOWLEDGEMENT

To whom it may concern. We would like to inform you that the data about the breast cancer patients were obtained according to the approval of the Central Research Committee Basra Health Directorate pursuant to the letter numbered (474) on 21/1/2020 from the Specialized Oncology Center in Basrah.

REFERENCES

Adi Pratama, F. R., & Oktora, S. I. (2023). Synthetic Minority Over-sampling Technique (SMOTE) for handling imbalanced data in poverty classification. *Statistical Journal of the IAOS*, vol. 39(1), pp. 233-239. doi:10.3233/SJI-220080

Afrash, M. R., Bayani, A., Shanbehzadeh, M., Bahadori, M., & Kazemi-Arpanahi, H. (2022). Developing the Breast Cancer Risk Prediction System using Hybrid Machine Learning Algorithms, *Journal of Education and Health Promotion*, vol. 11(1), pp. 1-12. doi:10.4103/jehp.jehp_42_22

Al-Hashimi, M. M. Y. (2021). Trends in Breast Cancer Incidence in Iraq During the Period 2000-2019. *Asian Pacific Journal of Cancer Prevention*, vol. 22(12), pp. 3889-3896. doi:10.31557/APJCP.2021.22.12.3889

Alotaibi, B. S., Alghamdi, R., Aljaman, S., Hariri, R. A., Althunayyan, L. S., AlSenan, B. F., Alnemer, A. M. (2024). The Accuracy of Breast Cancer Diagnostic Tools, *Cureus*, vol. 16(1), pp. 1-9. doi:10.7759/cureus. 51776

Al-Rikabi, A., & Husain, S. (2012). Increasing Prevalence of Breast Cancer among Saudi Patients Attending a Tertiary Referral Hospital: A Retrospective Epidemiologic Study, *Croatian Medical Journal*, vol. 53(3), pp. 239-243. doi:10.3325/cmj.2012.53.239

Asif, S., Wenhui, Y., ur-Rehman, S., ul-ain, Q., Amjad, K., Yueyang, Y., Jinhai, S., & Awais, M. (2024).

Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision. *Archives of Computational Methods in Engineering*, 1-31. doi:10.1007/s11831-024-10148-w

Azamjah, N., Soltan-Zadeh, Y., & Zayeri, F. (2019). Global Trend of Breast Cancer Mortality Rate: A 25-Year Study. *Asian Pacific journal of cancer prevention: APJCP*, vol. 20(7), pp. 2015-2020. doi:10.31557/APJCP.2019.20.7.2015

Battineni, G., Chintalapudi, N., & Amenta, F. (2020). Performance Analysis of Different Machine Learning Algorithms in Breast Cancer Predictions, *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 6(23), pp. 1-7. doi:10.4108/eai.28-5-2020.166010

Chakkouch, M., Ertel, M., Mengad, A., & Amali, S. (2023). A Comparative Study of Machine Learning Techniques to Predict Types of Breast Cancer Recurrence, *International Journal of Advanced Computer Science and Applications*, vol. 14(5), pp. 296-302. doi:10.14569/IJACSA.2023.0140531

Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning, *Computational Intelligence and Neuroscience*, vol. 2023(1), pp. 1-9. doi:10.1155/2023/6530719

Chtouki, K., Rhanoui, M., Mikram, M., Yousfi, S., & Amazian, K. (2023). Supervised Machine Learning for Breast Cancer Risk Factors Analysis and Survival Prediction. In: Lazaar, M., En-Naimi, E.M., Zouhair, A., Al Achhab, M., Mahboub, O. (eds) Proceedings of the 6th International Conference on Big Data and Internet of Things. BDIoT 2022. Lecture Notes in Networks and Systems, vol. 625, pp. 59-71. Springer, Cham. doi:10.1007/978-3-031-28387-1_6

Cuthrell, K. M., & Tzenios, N. (2023). Breast Cancer: Updated and Deep Insights. *International Research Journal of Oncology*, vol. 6(1), pp. 104-118.

Daly, A.; Rolph, R.; Cutress, R. I, & Copson, E. R. (2021) A Review of Modifiable Risk Factors in Young Women for the Prevention of Breast Cancer, *Breast Cancer: Targets and Therapy*, vol. 13, pp. 241-257. doi: 10.2147/BCTT.S268401

Dar, R. A., Rasool, M., & Assad, A. (2022). Breast Cancer Detection using Deep Learning: Datasets, Methods, and Challenges Ahead. *Computers in biology and medicine,* vol.149, pp. 1-23. doi: 10.1016/j.compbiomed.2022.106073

Darwich, M., & Bayoumi, M. (2024). An Evaluation of the Effectiveness of Machine Learning Prediction Models in Assessing Breast Cancer Risk, *Informatics in Medicine Unlocked*, vol. 49, pp. 1-17. doi:10.1016/j.imu.2024.101550

Dianati-Nasab, M., Salimifard, K., Mohammadi, R., Saadatmand, S., Fararouei, M., Hosseini, K. S., Jiavid-Sharifi, B., Chaussalet, T., & Dehdar, S. (2024). Machine Learning Algorithms to Uncover Risk Factors of Breast Cancer: Insights from a Large Case-Control Study, *Frontiers in Oncology*, vol. 13, pp. 1-13. doi:10.3389/fonc.2023.1276232

Edwards, T. L., Greene, C. A., Piekos, J. A., Hellwege, J. N., Hampton, G., Jasper, E. A., & Velez Edwards, D. R. (2023). Challenges and Opportunities for Data Science in Women's Health. *Annual Review of Biomedical Data Science*, vol. 6(1), pp. 23-45. doi:10.1146/annurev-biodatasci-020722-105958

Ferroni, P., Roselli, M., Buonomo, O., Spila, A., Portarena, I., Laudisi, A., Valente, M., Pirillo, S., Fortunato, L., Costarelli, L., Cavaliere, F., & Guadagni, F. (2018). *Anticancer Research*, vol. 38(8), pp. 4705-4712. doi: 10.21873/anticanres.12777

Garcia-Moreno, F. M., Ruiz-Espigares, J., Gutiérrez-Naranjo, M. A., & Marchal, J. A. (2024). Using Deep Learning for Predicting the Dynamic Evolution of Breast Cancer Migration, *Computers in Biology and Medicine*, vol. 180, pp. 1-18. doi:10.1016/j.compbiomed.2024.108890

González-Castro, L., Chávez, M., Duflot, P., Bleret, V., Martin, A. G., Zobel, M., Nateqi, J., Lin, S., Pazos-Arias, J.J., Del Fiol, G., & López-Nores, M. (2023). Machine Learning Algorithms to Predict Breast Cancer Recurrence using Structured and Unstructured Sources from Electronic Health Records, *Cancers*, vol. 15(10), pp. 1-16. doi:10.3390/cancers15102741

Iparraguirre-Villanueva, O., Epifanía-Huerta, A., Torres-Ceclén, C., Ruiz-Alvarado, J., & Cabanillas-Carbonel, M. (2023). Breast Cancer Prediction using Machine Learning Models, *International Journal of Advanced Computer Science and Applications*, vol. 14(2), pp. 610-620. doi:20.500.13053/9106

Jain, B., & Singla, N. (2023). Breast Cancer Detection using Machine Learning Algorithms. *Journal of Computers, Mechanical and Management*, vol. 2(6), pp. 30-35. doi: 10.57159/gadl.jcmm.2.6.230109

Lee, M. (2023). Deep Learning Techniques with Genomic Data in Cancer Prognosis: A Comprehensive Review of the 2021–2023 Literature, *Biology*, vol. 12(7), pp. 1-22. doi:10.3390/biology12070893

Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., Peng, X. (2021). Predicting Breast Cancer 5-Year Survival using Machine Learning: A Systematic Review, *PloS one*, vol. 16(4), pp. 1-23. doi: 10.1371/journal.pone.0250370

Manikandan, P., Durga, U., & Ponnuraja, C. (2023). An Integrative Machine Learning Framework for Classifying SEER Breast Cancer, *Scientific Reports*, vol. 13(1), pp. 1-12. doi:10.1038/s41598-023-32029-1

Mohaimenul, I., & Poly, T. N. (2019). Machine Learning Models of Breast Cancer Risk Prediction. *bioRxiv*, pp. 1-5. doi:10.1101/723304

Mohsin, R. N., & Mohamad, B. J. (2024). Clinical and Histopathological Features of Breast Cancer in Iraqi

Patients between 2018-2021. *Iraqi Journal of Science*, vol. 65(1), pp. 90-107. doi:10.24996/ijs.2024.65.1.9

Mueller, T., Segin, A., Weigand, C., & Schmitt, R. H. (2023). Feature Selection for Measurement Models. *International Journal of Quality & Reliability Management*, vol. 40(3), pp. 777-800. doi: 10.1108/ IJQRM-07-2021-0245

Naji, M. A., El Filali, S., Aarika, K., Benlahmar, El H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis, *Procedia Computer Science*, vol. 191, pp. 487-492. doi:10.1016/j.procs.2021.07.062

Obuchowicz, R., Strzelecki, M., & Piórkowski, A. (2024). Clinical Applications of Artificial Intelligence in Medical Imaging and Image Processing - A Review, *Cancers*, vol. 16(10), pp. 1-16. doi:10.3390/cancers16101870

Parekh, D. H., & Dahiya, V. I. S. H. A. L. (2023). Early Detection of Breast Cancer Using Machine Learning and Ensemble Techniques. *International Journal of Computing*, vol. 22(2), pp. 231-237. doi:10.47839/ ijc.22.2.3093

Poornajaf, M., & Yosefi, S. (2023). Improvement of the Performance of Machine Learning Algorithms in Predicting Breast Cancer. *Frontiers in Health Informatics*, vol. 12, pp. 1-7. doi: 10.30699/fhi. v12i1.400

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022). Prediction of Breast Cancer using Machine Learning Approaches, *Journal of Biomedical Physics & Engineering*, vol. 12(3), pp. 297-308. doi: 10.31661/jbpe.v0i0.2109-1403

Reshan, M. S. A., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., Azar, A. T., & Shaikh, A. (2023). Enhancing Breast Cancer Detection and Classification using Advanced Multi-Model Features and Ensemble Machine Learning Techniques, *Life*, vol. 13(10), pp. 1-20. doi:10.3390/life13102093

Roheel, A., Khan, A., Anwar, F., Akbar, Z., Akhtar, M. F., Imran Khan, M., Sohail, M.F., & Ahmad, R. (2023). Global Epidemiology of Breast Cancer Based on Risk Factors: A Systematic Review. *Frontiers in Oncology*, vol. 13, pp. 1-15. doi: 10.3389/fonc.2023. 1240098.

Syamsiah Mashohor, D. N. F. P. M., Mahmud, R., Hanafi, M., & Bahari, N. (2023). Transition of Traditional Method to Deep Learning Based Computer-Aided System for Breast Cancer using Automated Breast Ultrasound System (ABUS) Images: A Review. *Artificial Intelligence Review*, vol. 56, pp. 15271-15300. doi:10.1007/s10462-023-10511-6

Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., Poorolajal, J. (2019). Prediction of Survival and Metastasis in Breast Cancer Patients using Machine Learning Classifiers, *Clinical Epidemiology and Global Health*, vol. 7(3), pp. 293-299. doi:10.1016/j.cegh.2018.10.003

Yadav, R. K., Singh, P., & Kashtriya, P. (2023). Diagnosis of Breast Cancer using Machine Learning Techniques - A Survey. *Procedia Computer Science*, vol. 218, pp. 1434-1443. doi:10.1016/j.procs.2023.01.122

Zhang, S., Jin, Z., Bao, L., & Shu, P. (2024). The Global Burden of Breast Cancer in Women from 1990 to 2030: Assessment and Projection Based on the Global Burden of Disease Study 2019. *Frontiers in Oncology*, vol. 14, pp. 1-13. doi: 10.3389/fonc.2024.1364397

Zhu, J. J., Yang, M., & Ren, Z. J. (2023). Machine Learning in Environmental Research: Common Pitfalls and Best Practices, *Environmental Science & Technology*, vol. *57*(46), pp. 17671-17689. doi:10.1021/acs.est.3c00026

BIOGRAPHIES

Salma Abdulbaki Mahmood received the B.S. degree in Computer Science from University of Basrah, Basrah, Iraq, in 1988. She then pursued an M.S. degree in Computer Science/ Artificial Intelligence from University of Basrah, graduating in 1993. She completed her Ph.D. in Computer Science/Artificial Intelligence with a dissertation on Knowledge Based Lexicon for Arabic Language Understanding at University of Basrah, Basrah, Iraq, in 2000. Currently, Dr. Mahmood is an Assistant Professor at the Intelligent Medical Systems Department of Computer Science and Information Technology Collage, University of Basrah, where she teaches Artificial Intelligence, Machine Learning, Natural Language Processing and Programming Language, and conducts research in the areas of AI, ML and NLP. Her work primarily focuses on ML with NLP and medical applications. Prior to her current position, she held the position as the dean of Computer Science and Information Technology College, University of Basrah in (2019-2023). Her current research projects involve creating medical prediction systems, aiming to improve medical services in her country and developing Arabic language applications.

Myssar Jabbar Hammood Al-Battbootti received a B.S. degree in Computer Engineering from the Iraq University Basra, Iraq, in 2015 and an M.S. degree in Software Engineering from the National University of Science and Technology Politehnica Bucharest, Romania, in 2020. He is currently studying for a PhD degree in Computer Engineering Technology at the National University of Science and Technology Politehnica Bucharest, Romania. Mr. Al-Battbootti has expertise in machine learning, operating systems, computer networks, software engineering, distributed systems, technologies for big data analysis, and Internet of Things.

Saad Shaheen Hamadi Al-Taher received the B.S. degree in Medicine and Surgery, College of Medicine, from University of Basrah, Basrah, Iraq, in 1986. He then pursued Board of Internal Medicine Gastroenterology Iraqi Board of Medical Specialties Iraq in 1992. Currently, Dr. Al-Tahir is a Professor at the Internal Medical Department of Medicine Collage, University of Basrah, where he teaches Internal Diseases and Treatments, and he is member of the exams committee of Arabic and Iraqi board exam for medical specialty. He conducts research in the areas of GIT and infectious diseases. His work primarily focuses on gastrointestinal diseases and their infections. He is a member of multiple Medical Societies, namely G.I.T. Society, Diabetes Society, Iraqi Medical Association Arab Endocrine Society Patents. Prior to his current position, he held position as Chancellor of University of Basrah, Basrah, Iraq, between 2019-2024. Currently, he is improving teaching and medical services.

Iuliana Marin received the B.S. degree in computer science, the M.S. degree in software engineering, and the Ph.D. degree in computers and information technology from the National University of Science and Technology Politehnica Bucharest, Romania, in 2015 and 2017, respectively. She is currently an Associate Professor with the University of Science and Technology Politehnica Bucharest and she has expertise in databases, operating systems, computer networks, the semantic web, distributed systems, technologies for big data analysis, and the Internet of Things. She received several prizes, including the Gold Medal for the invention “Preeclampsia and detection of secondary diseases based on heart rate and a weighted stratified network” at the Salon of Scientific Research, Innovation and Inventions “PRO INVENT,” Cluj-Napoca, Romania, in 2020, and the Silver Medal for the invention “Automatic system for establishing the diet in case of preeclampsia” at the International Exhibition of Inventions and Innovations “Traian Vuia,” Timişoara, Romania, in 2019.

Costin-Anton Boiangiu was born in Bucharest, Romania, on July 28, 1972. He received his engineer degree in computer science from the Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Romania, in 1995. He completed his Advanced Studies in Computer Systems Architecture at the same institution in 1996, followed by a Ph.D. in computer science in 2003. His doctoral research focused on methods for generating multimedia applications and virtual reality environments. He began his academic career as a Teaching Assistant at the University Politehnica of Bucharest in 1995, progressing to Lecturer in 1998, Senior Lecturer in 2003, Associate Professor in 2009, and finally Professor in 2016 in the Department of Computer Science. His professional experience also includes collaborations with various companies like Electronic Arts, CCS-Content Conversion Specialists GmbH, and UTI Group, where he served as a software developer and project leader. His research interests include real-time 3D scene visualization, multimedia applications, and software project management. He has authored over 100 scientific publications, including 10 books. Prof. Boiangiu is a Senior Member of the IEEE and a member of several other professional societies. He has received numerous awards for his contributions to computer science education and research. He has also served as a reviewer for various IEEE conferences and journals and has been involved in organizing several international conferences and scientific events.

Nicolae Goga is a professor at the National University of Science and Technology Politehnica Bucharest and a part-time senior researcher at the Molecular Dynamics Group at the University of Groningen in the Netherlands. Mr. Goga obtained his first PhD in 2004 with the thesis "Control and Selection Techniques for Automated Testing of Reactive Systems". The second PhD was obtained in 2005 with the thesis "Contributions to the modeling and verification of reactive systems". In 2012, he received the IEEE Technology Merit Award. Mr. Goga has been an active member of several research projects. The algorithms developed in the molecular dynamics project are part of the Gromacs molecular dynamics package and are used by prestigious groups worldwide at Stanford University, Cambridge, and powerful companies like Intel and IBM. Other research projects where he was project director: Eurostar Project i-Bracelet, Eurostar Project i-Light, Eureka Project Questor, Improved parallel scaling of Gromacs MD code, Improving the quality of protocol standards. His expertise is broad, including IoT, artificial intelligence, distributed systems, molecular dynamics, graphics, animation, and networking.