Journal of Computer Science

Original Research Paper

Development of Big Data Classifier for Biomedicine Early Diagnosis: An Experimental Approach Using Machine Learning Methods

[1]Ma Beth Solas Concepcion, [2]Bobby Dioquino Gerardo, [3]Frank Ibañez Elijorde, [4]Joel Traifalgar De Castro and [5]Nerilou Bermudez Dela Cruz

[1]Department of Information Systems, Faculty of College of Information and Communications Technology, West Visayas State University, La Paz, Iloilo City, Philippines
[2]Division of Computer Science, Faculty of College of Information and Communications Technology, West Visayas State University, La Paz, Iloilo City, Philippines
[3]Division of Information Technology, Faculty of College of Information and Communications Technology, West Visayas State University, La Paz, Iloilo City, Philippines
[4]Division of Entertainment, Multimedia and Computing, Faculty of College of Information and Communications Technology, West Visayas State University, La Paz, Iloilo City, Philippines
[5]Division of Computer Science, Faculty of College of Information and Communications Technology, West Visayas State University, La Paz, Iloilo City, Philippines

Article history
Received: 09-09-2023
Revised: 20-11-2023
Accepted: 28-11-2023

Corresponding Author:
Ma Beth Solas Concepcion
Department of Information Systems, Faculty of College of Information and Communications Technology, West Visayas State University, La Paz, Iloilo City, Philippines
Email: mbconcepcion@wvsu.edu.ph

Abstract: In the fast-phase world, data availability is abundant due to a rapid adaptation increase of big data technologies. Large amounts of data have been generated and collected at an unprecedented speed and scale, introducing a revolution in medical research practices for biomedicine informatics. Thus, there is an immense demand for statistically rigorous approaches, especially in the medical diagnosis discipline. Therefore, this study utilized the Bayesian Belief Network (BBN) for feature selection, which identifies relevant features from a larger set of attributes and employs a stratification for the Stochastic Gradient Descent (SGD) classifier in the classifying of breast cancer on the publicly available machine learning repository at the University of California, Irvine (UCI) such, breast cancer Wisconsin and Coimbra breast cancer datasets. The experimental approach of using BBN as feature selection achieved 0.95% coincidence. Thus, a stratified Stochastic Gradient Descent (SGD) was employed to build a classification model to validate the coincidence. Our proposed modeling classifier approach reached novelty 98%, which improved by 7% compared to the previous works. Furthermore, this study presents a web-based application, a prototype type, to employ the proposed classifier model for breast cancer diagnosis. This study expects to provide a source of confidence and satisfaction for medical physicians to use decision-support tools.

Keywords: Biomedicine Diagnosis Application, Big Data, Data Mining, Classification Algorithm, Bayesian Belief Network

Introduction

The increasing success of Machine Learning (ML) as a tool for discovering knowledge from a large and diverse dataset or Big Data is evident with its ability to uncover underlying structures and patterns and make accurate predictions used in various industries, especially healthcare (Alfian *et al*., 2022). Presently, there is a continuous widespread adoption of Machine learning ML in the healthcare domain by leveraging in analyzing large amounts of patient data. ML algorithms detect patterns and trends not immediately apparent to human observers, allowing healthcare providers to develop more targeted and effective treatment plans with these advances in new technologies and advancements, an evolving field of medicine continues. Medical physicians are constantly

looking for new ways to improve diagnostic procedures and treatments for their patients by enabling early interventions to prevent serious complications and improve patient outcomes (Naqvi *et al*., 2021).

Thus, an early diagnosis is crucial for effectively managing various health conditions. For instance, early diagnosis is the most effective and efficient way to use as a tool for decision making which provides aid for medical practitioners. In fact, one of the most prevalent cancers worldwide is breast cancer and a top cause of death in women worldwide is 12.5% of all newly diagnosed cancer cases yearly (CDC, 2022; Curtin *et al*., 2023). There are two types of breast tumors, which are categorized to: Benign and malignant. The benign tumors do not pose a threat to the body and do not spread to other areas of the body. They can be removed surgically and do not return. The malignant tumors, on the other hand, are extremely dangerous and can metastasize to other parts of the body. Even if surgically removed, malignant tumors will likely recur (Koscielny, 1999; Wardle *et al*., 2015).

Furthermore, a woman's chances of developing breast cancer increase by twofold one, if she has a first-degree relative, such as her mother, sister, or daughter, who has been diagnosed with breast cancer disease. Second, about fifteen percent of women in a family members have also been diagnosed with breast cancer (Sivakami and Saraswathi, 2015) Thus, this study will develop an early detection model for breast cancer using an ML algorithm for classifying two categories, benign and malignant. With this, by detecting symptomatic patients as early as possible, healthcare professionals can design tailored treatment plans that improve survival rates, decrease complications, and lessen the overall cost of care (Akben *et al*., 2019).

Numerous studies have been using ML algorithms for the early diagnosis or detection of breast cancer and have shown significant performance (Nanglia *et al*., 2022). However, in this study, the proponents utilized specifically the Bayesian Belief Network (BBN) for classifying breast cancer, which is fitting for its "symbolic representation and handling of uncertainty, where different scenarios are possible by given evidence (Kharya *et al*., 2014)".

Kharya *et al*. (2014) proposed "Bayesian belief networks for prognosis and diagnosis of breast cancer," which showed the utilization of BBN as an automated detection support tool for breast cancer authors showed practical modeling and analysis approaches that are based on the widely studied Wisconsin Breast Cancer Database (WCBD). Another study that similarly implemented this algorithm, such as the Weighted Bayesian belief network from Kharya *et al*. (2022), designed and developed an automated model using the Bayesian Belief Network (BBN) for constructing expert knowledge diagnosis. An expert knowledge diagnosis performs complex data processing which is similar to the evaluation made by human experts. The authors obtained strong rules from attributes and class labels by applying automated weight assignment using the open Markov tool. The dataset was partitioned using the minimum threshold method, which the authors defined and categorized. Based on different thresholds, the proposed predictive modeling results got the highest, 97.18% (Kharya *et al*., 2014; 2022).

Dag *et al*. (2022) developed a decision support tool, a "tree-augmented Bayesian belief network," for big data-driven breast cancer survival prediction and understanding. The authors utilized a probabilistic graphical model using BBN in a tree-augment manner to understand the underlying factors and interrelations affecting breast cancer. The TAN is a method to structure the variables affecting cancer. The authors also integrated a Genetic Algorithm (GA) to calibrate the parameters selection and balance the data, a common problem in a significant data domain. The GA integration with the TAN method identified the hidden conditional probabilistic dependencies among the breast cancer factors. The results of this study used What-if analysis and the model could predict survival with 0.813 accuracies (Dag *et al*., 2022).

However, first, none of these proposed studies employed stratification for feature selection; for instance, the study Ong (2011) proposed to improve the Bayesian belief network using structural learning and found that learning the dependencies among the data features and constructing the added relationship when working on classification models helps to improve the accuracy of the Naïve Bayes (NB) classification, specifically when the dependency is highly strong. Second, they were not able to integrate their model web-based application for the viability of their proposed models (Ong, 2011). Thus, these mentioned studies will serve as our benchmark and comparison. In addition, the proponents integrate the proposed model into a web-based application which could help medical physicians in their decision-making process and provide preventive actions for their patients that can be taken earlier than the incidents occur. Therefore, the proponents' contributions to the present study are as follows:

- Utilized stratified BBN to identify useful feature selection breast cancer dataset
- Employing stratified Stochastic Gradient Descent (SGD) classifier to classify breast cancer dataset
- Undertook in-depth experiments and analyses comparing of using stratification in feature selection to other prediction models and findings from previous works

- Finally, the proponents created a web-based application for the breast cancer dataset, a classifier tool to illustrate the feasibility and viability of the model

Materials and Methods

Dataset

In this study, the proponents used the breast cancer dataset from two primary sources, WCBD and Coimbra, which shows the instances, classes, and number of attributes in Table 1, publicly available at the University of California Irvine (UCI), a machine learning repository. First, Wisconsin diagnostic and prognostic breast cancer from Dr. William H. Wolberg's clinical work at the University of Wisconsin Hospitals in Madison. As of 1992, Dr. Wolberg had collected 699 patient diagnoses and, as of 1984, had collected 198 instances of patient prognostic, in which the numerical attributes are computed from digitized images of fine needle aspirate (fna) for a breast mass that defines the features of the cell nuclei present in the image (Wolberg *et al.*, 1995). Second, the Coimbra Breast Cancer Dataset (CBCD) from the Department of Gynaecology at the University Hospital Centre of Coimbra between 2009 and 2013 consists of sixty-four women who have breast cancer and 52 healthy subjects. The Coimbra Breast Cancer Dataset (CBCD) was composed of routine blood analyses mainly, such as glucose, insulin, HOMA, leptin, adiponectin, resistin, MCP-1, age, and Body Mass Index (BMI) to predict the presence of breast cancer (Patricio *et al*., 2018).

Modeling

The proposed model consists of a Bayesian belief network for feature selection and stratification to classify breast cancer shown in Fig. 1. First, the proponents employed preprocessing, which involves the following process methods cleaning, normalization, and stratification.

Cleaning the dataset is necessary to remove the null and empty values; after the dataset is cleaned, the proponents utilize normalization to rescale the attribute's class string values into numerical values (0, 1). Moreover, the last method for preprocessing is stratification. The purpose of employing stratification as one of the problems that the proponents encountered in early diagnostic systems was the multiplicity of features such as the irrelevancy and redundancy in these features which increases the confusion of the classification model and decreases learning precision (Aalaei *et al*., 2016). Feature selection is one of the methods that can cope with this problem, it plays a vital role in classification and is extensively used in statistics, pattern recognition, and the medical domain (Aamir *et al*., 2022). Ong (2011) found that learning the dependencies among the features when doing classification models helps to improve the accuracy of the NB classification by using stratification (Ong, 2011).

The feature selection using the Bayesian belief network using bnlearn a Python package. The design of the unlearn package is to facilitate this learning process with probabilistic models that represent the dependencies among a set of variables using a Directed Acyclic Graph (DAG). Learning the graphical structure of Bayesian networks is fundamental in building effective models for decision-making, risk assessment, and inference. The bnlearn package is an automated learning of the graphical structure of Bayesian networks, which involves discovering the relationships and dependencies between variables, which is essential for modeling complex systems (Peng and Ding, 2012). Furthermore, in this study, the proponents utilized stratified ten-fold cross-validation or K-fold CV to evaluate the accuracy of the proposed BBN classifier model. The average values computed in the loop are the performance measures reported by k-fold cross-validation (Lundervold et al., 2019). This approach can be computationally expensive but only uses a little data, which is highly significant in problems such as inverse inference where the number of samples is minimal.

Table 2 shows the confusion matrix, a helpful tool for analyzing and representing the classification model outline in matrix form. It shows how classifier performances are correct and incorrect per class where TP here stands for true positive predictions, for instance, in a binary classification such as classifying benign cancer as 1, TP gives the count of the number of 1 s correctly classified as 1, i.e., the number of benign cancer classified as benign. TN represents true negative predictions, i.e., number of 0 s, malignant cancer, classified as 0. False Positive (FP) is the count of the number of malignant cancers classified as benign and False Negative (FN) is the count of benign cancers classified as malignant. Moreover, this study also utilizes average performance metrics such as accuracy (%), sensitivity (%), and Receiving Operating Characteristics (ROC) curves by conducting 10 runs of stratified 10-fold CV.

Table 1: Breast cancer dataset

Dataset	Attributes	Instances	Classes
Original dataset BC Wisconsin	11	699	
Diagnostic BC Wisconsin	32	569	2
Coimbra BC	9	76	2

Table 2: BBN classifier confusion matrix

		Predicted class	
		1	0
Actual class	1	TP	FN
	0	FP	TN

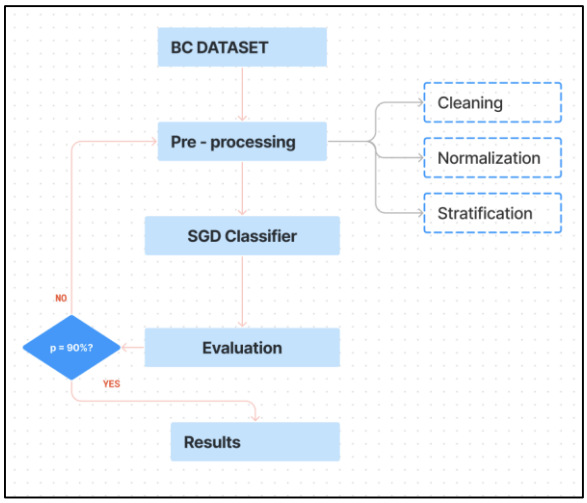

Fig. 1: The proposed flow diagram for BBN classification modeling for breast cancer

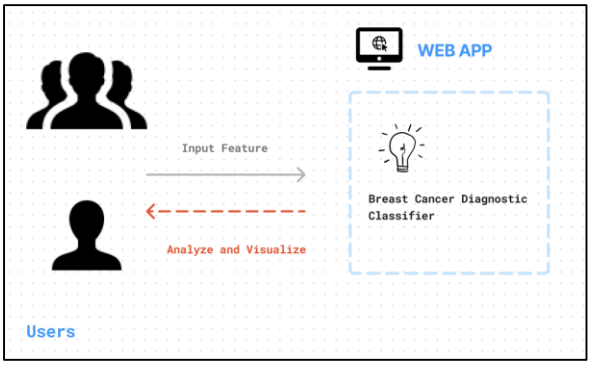

Fig. 2: The proposed designed framework for a web-based breast cancer classifier

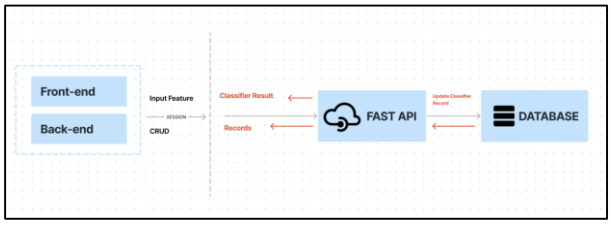

Fig. 3: The proposed flow diagram for back and front-end application development

Materials

The primary materials for developing the proposed web application are Python development using fast API for restful implementation, MongoDB for data storage, and front scripting, including HTML, JavaScript, and CSS framework Tailwind. The proponents implemented the ML model using Python Script's latest version using several built-in frameworks such as Numpy, CSV, and Pandas. The experiments were performed on a Ryzen 9 laptop with 16 GB RAM running on Windows 11 64-bit. The impact of the Python version on model training is multifaceted, encompassing compatibility with libraries, performance improvements, GPU acceleration, bug fixes, ecosystem support, and community engagement. Thus, to optimize the model training process, selecting a Python version that aligns with the requirements of the machine learning stack while considering the benefits of the latest enhancements and features is advisable. The proponents ensured that the version of all requirements was compatible with the device used, enabling more efficient utilization of multiple cores during training, improving execution speed and resource utilization, and impacting the overall training time, especially for computationally intensive modeling. Furthermore, the proponents used a Ryzen 9 Laptop with 16 GB RAM running on Windows 11 64-bit, which the type and power of the device, especially the Graphics Processing Unit (GPU), significantly impact the speed at which complex mathematical operations, such as matrix multiplications in neural networks, are executed. A device with high computational power can dramatically reduce the time required for model training. The memory capacity of a device influences the size of datasets and models during training. Larger memory capacity is crucial for handling extensive datasets and deep neural network architectures without encountering memory limitations that could impede the training process. The cost associated with acquiring and maintaining computing devices is a practical consideration. Balancing performance requirements with budget constraints is essential to ensure cost-effective model training without compromising the trained models' quality.

Implementation

In this study, after the model's evaluation, it will be incorporated into two web applications, such as front-end and back-end implementation shown in Fig. 3. In front-end integration, the users input the x-attributes and the server will process it and provide the results shown in

Figure 2 on the other hand, the back-end integration is a prototype application for the feasibility of how the entire system would work. Therefore, the back-end integration has admin tasks such as creating, reading, updating, deleting the data, and authenticating whoever accesses the system. The web application components, for instance, in the front-end, the dashboard and classifier results, and for the back-end, the authentication of login/logout, the crud components such as a physician, patient, diagnostic test, and lastly, classification results.

Results and Discussion

In this section, the proponents assess how the proposed machine learning model performed and how feature

selection affected the model's accuracy. This section is divided into two sections: (a) Model's performance and (b) Prototype implementation.

Model's Performance

The model's performance measures how accurately it can achieve its intended task for a particular problem, such as in this proposed study, handling big data. Thus, this section is divided into two sub-sections: First, the feature selection results using the BBN graphical model. Second, the stratified stochastic gradient descent classifier results.

Feature Selection

The proponents used the BBNs for feature selection, which is choosing a subset of significant features from a larger set of data attributes. Constructing a BBN allows you to assess the dependencies and conditional relationships between attributes. Features that are strongly connected to the target variable or other important attributes can be considered more relevant. Thus, BBNs help identify dependencies between attributes to analyze the network's conditional probability distributions and understand how attributes influence each other. Attributes with strong conditional dependencies on other attributes or the target variable can be considered important features.

Figure 4 shows the BBN feature selection results, and the strongly connected attributes based on the target variable, 'diagnosis'. Based on the results, the strongly dependent variables are 'radius_mean', 'texture_mean', 'perimeter_mean', 'radius_se', 'perimeter_se', 'area_se', 'radius_worst', and 'perimeter_worst'. The attribute network shows coincidence ≈ 0.95 , which these attributes highly influences each other based on the target variable, 'diagnosis'. Therefore, this implies two parts: First, there is a relationship or dependency between these attributes. Second, changes or variations can affect or correlate with changes in the relationship or dependency between these attributes. Moreover, Fig. 6 shows how highly correlated the features are. The correlation of these attributes on each is particularly significant considering the target variable, which is crucial in the classification accuracy of the SGD classifier. In addition, Fig. 5 shows the feature selection inference using the Coimbra breast cancer dataset, which shows only two feature attributes had relevant features and achieved coincidence ≈ 0.5; therefore, in this study, the researchers focus on using the Wisconsin breast cancer dataset as the attributes indicate relevant features.

Classifier

This section shows the proposed model's performance, which refers to how effectively a model can make classifications on new, unseen data.

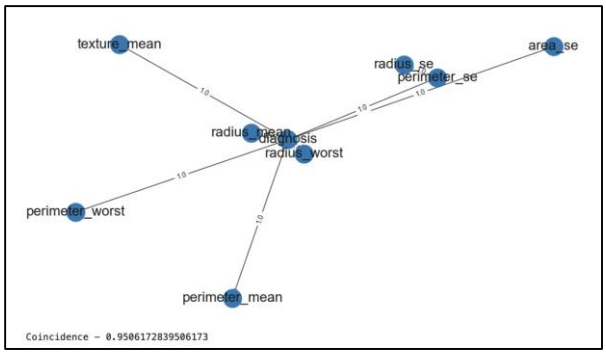

Fig. 4: BBN feature selection results for diagnostic BC Wisconsin

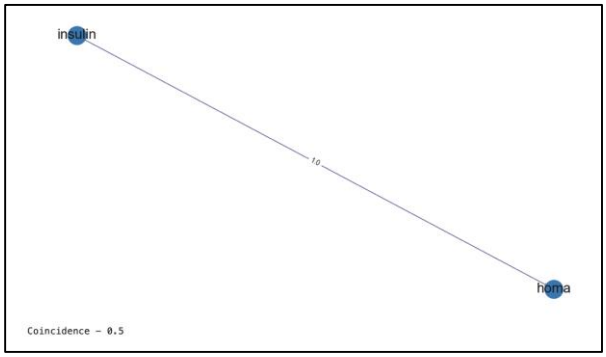

Fig. 5: BBN feature selection results for Coimbra BC

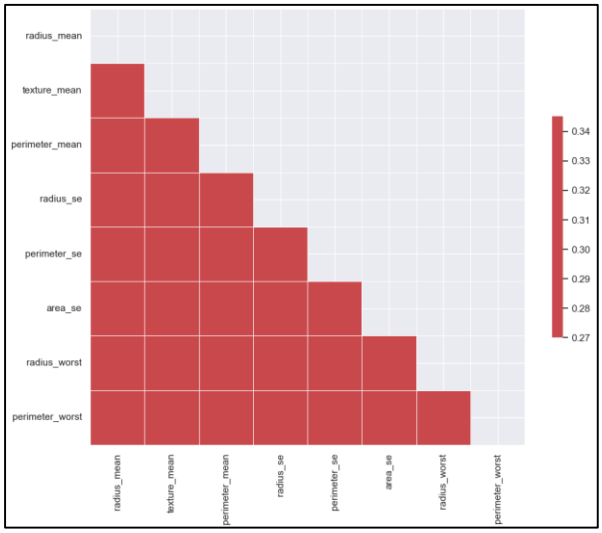

Fig. 6: Attribute correlation from BBN feature selection results for diagnostic BC Wisconsin

The proponents proposed a classifier model, the SGD classifier, to classify the breast cancer datasets and evaluate and compare them to other data-driven models, namely, Bernoulli Naive Bayes, Gaussian Naive Bayes, AdaBoost and ensemble with the three Bayesian Models and including the proposed model. The first evaluation uses the confusion matrix, a detailed breakdown of the

performance of the proposed classifier by summarizing the predictions on a set of data, which allows for the calculation of the various performance metrics, which includes accuracy, precision, recall, and F1-score and can be seen more detailed in Table 3.

Figures 7-8 show how the proposed SGD classifier works, simulating the test dataset, which views how well a classification model performs, considering correct and incorrect predictions for each class. Figures 9-10 show the proposed model, SGD results for Receiver Operating Characteristic (ROC), and 2-class Precision-Recall curve, these evaluation metrics indicate the false positive and false negative results contrast to specify an unbalanced dataset. The results suggest that the proposed model had ≈ 0.87 and based on (Huang and Ling, 2005), AUC ≈ 1 is considered the best model. Therefore, the proposed model is closer to the target AUC ≈1. In the area under the curve and receiver operating characteristic (AUC-ROC) which represents the overall performance of the classification model. A higher AUC-ROC value (closer to 1) indicates better discrimination between positive and negative instances. ROC curves visually represent how well a model can distinguish between two classes across different decision thresholds. The model assigned a probability or score to each instance in binary classification and set a threshold value to determine the classification. The ROC curve is created by varying this threshold value and calculating each threshold's sensitivity and specificity values. Therefore, the proposed model's evaluation suggests a strong ability that discriminate between the positive and negative classes. However, the proponents see room for improvement in future studies.

Table 3: Performance metrics for classification

Method	Accuracy (%)	Precession (%)	Sensitivity/ recall (%)	F1-score (%)
Proposed model (SGD classifier)	98.36	99.00	99.00	99.00
Bernoulli NB	57.97	34.10	57.97	73.01
Gaussian NB	57.97	34.10	57.97	73.01
Ensemble	95.74	95.89	95.74	95.77
Adaboost	96.27	96.30	96.27	96.28

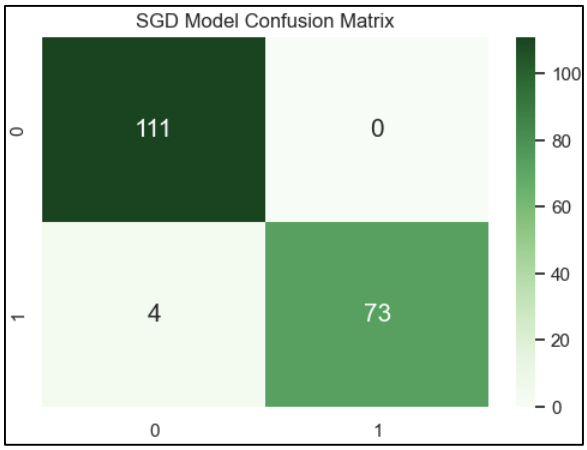

Fig. 7: Confusion matrix stratified structure stochastic

Fig. 8: Class prediction stratified structure stochastic gradient descent classifier

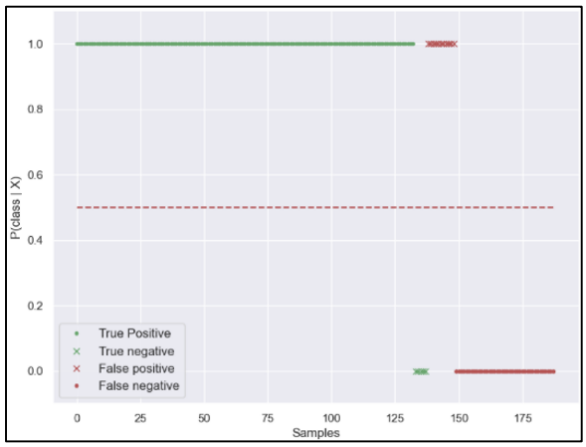

Fig. 9: Stratified stochastic gradient descent classifier

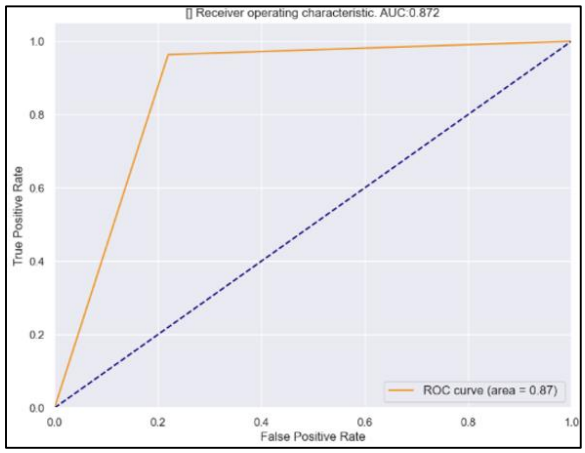

Fig. 10: Stratified Stochastic gradient descent classifier results for Receiver Operating Characteristic (ROC)

Prototype Implementation

This section previews the web-based application prototype implementation for the proposed model,

which involves building a simplified, interactive version to demonstrate its core features and functionalities, which is to be able to classify a breast cancer dataset by integrating the proposed model. Implementing web-based diagnostic and screening tools has been widely utilized by researchers, for instance, Yu *et al*. (2020) presented a healthcare assessment for liver diseases, and our study's benchmark from Alfian *et al*. (2022) a diagnostic screening for breast cancer using Coimbra dataset to facilitate preliminary diagnostics to aid initial screenings for patients. Therefore, one of the proponent's objectives in this study is to visualize and test whether the proposed model is feasible for a diagnostic and screening tool. This prototype implements a preliminary web application composed of a landing page, dashboard components, and two experimental interfaces for Winsconsin and Coimbra datasets. Figure 11 shows the landing page where the user can sign in and use the core components of the proposed prototype. Figure 12 shows the components of the web-based application, such as the dashboard page, diagnostic page, patient page, physician page, and settings. Figures 13-14 show the interface form of two diagnostic sub-pages, such as for the Winconsins and Coimbra datasets, and have the same process flow. Figure 13 shows the input features, which are required input, 'radius_mean', 'texture_mean', 'perimeter_mean', 'radius_se', 'perimeter_se', 'area_se', 'radius_worst' and 'perimeter_worst'. When all the fields have been filled out, the users can click the "submit" button to send the information needed to a secure remote server, which loads the proposed model and processes the new input dataset. In addition, the patient diagnosis ID label is the only identifier and the patient selection is tokenized to anonymize the patient's privacy. Hence, this prototype web-based application aims to assist individuals by offering early diagnoses for preventive measures or further analysis of results. The proposed classifier's efficiency and real-time data processing capabilities make it well-suited for scenarios requiring rapid predictions, particularly in clinical settings, which can contribute to quicker decision-making and interventions. The SGD classifier is adept at handling extensive datasets and is easily scalable, a crucial aspect in clinical applications with diverse information from numerous patients and various medical parameters. Additionally, the stochastic nature of SGD enables online learning, allowing the model to adapt to evolving patient data over time, which proves advantageous in dynamic clinical environments where patient conditions and healthcare practices change.

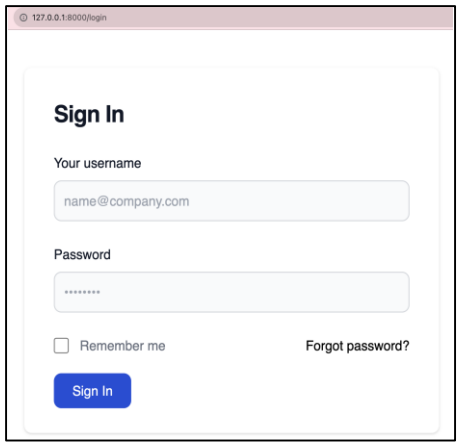

Fig. 11: Landing page web-based prototype application

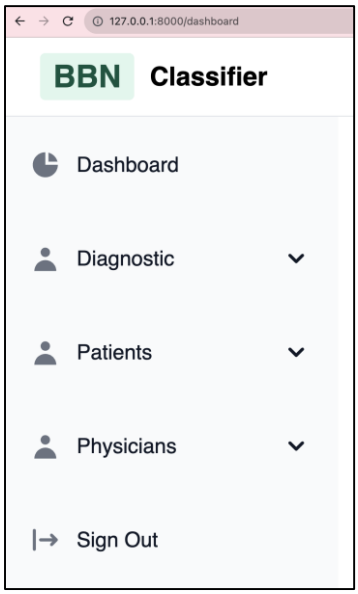

Fig. 12: Dashboard components web-based prototype application

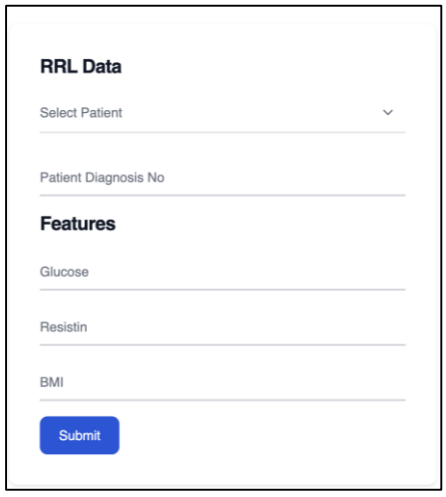

Fig. 13: Input interface for Wisconsin diagnostic dataset web-based prototype application

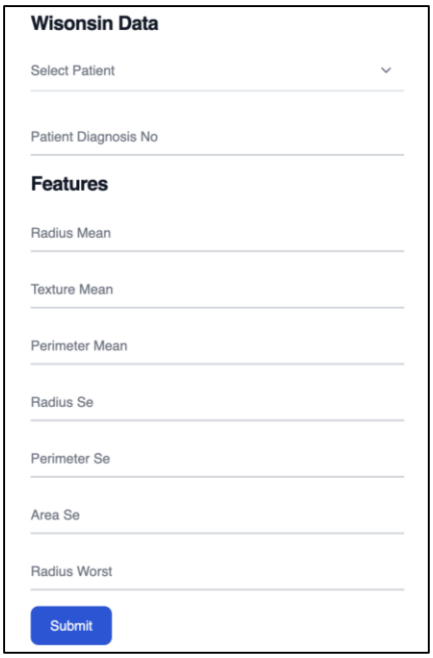

Fig. 14: Input interface for Coimbra dataset web-based prototype application

Table 4: Performance from previous works

Author/method	Feature selection	Accuracy (%)	Practical application
Proposed model (SGD)	BBN	98.00	Yes
Rasool *et al*. (2022)	RFE	76.42	No
Polynomial model Alfian *et al*. (2022)	Extra-trees	80.23	Yes
SVM and extra-trees Kharya *et al*. (2014)/BBN	Construction algorithm and ensemble	90-91.83	No

Notes: Recursive Features Elimination (RFE), Bayesian Belief Network (BBN), Stochastic Gradient Descent (SGD)

Table 3 shows the proposed model performance with other data-driven models, such as Bernoulli Naive Bayes, Gaussian Naive Bayes, AdaBoost, and ensemble, combining the three Bayesian models, including the proposed model SGD. The results indicate the proposed model achieved 98.36% average accuracy compared to other data models. The proposed model processes training examples one at a time, making it computationally efficient and memory-friendly, leading to faster convergence than batch gradient descent. Significantly, the use of feature selection helps prevent overfitting by focusing on the most relevant features, such as 'radius_mean', 'texture_mean', 'perimeter_mean', 'radius_se', 'perimeter_se', 'area_se', 'radius_worst' and 'perimeter_worst.' A subset of features reduces the computational resources required for training and making predictions. In addition, the AdaBoost and ensemble models also achieved good average results, with a difference ranging from 2.62-2.09% behind the proposed model. Additionally, Table 4 provides a comparative analysis with previous research in general. The studies by Rasool *et al*. (2022); Alfian *et al*. (2022) closely resemble our proposed study. However, our model demonstrated a notably high average accuracy of 98.36%. In contrast, Alfian *et al*. (2022) presented an SVM and Extra-trees with web-based applications, achieving an average accuracy of 80.23%. Consequently, our proposed model, embedded in a web-based prototype application, emerges as a robust approach in clinical settings, surpassing the studies by Rasool *et al*. (2022); Alfian *et al*. (2022).

Conclusion

The proponents proposed a classifier model based on the Stochastic Gradient Descent (SGD) and Bayesian Belief Network (BBN) as a feature selector using simulation-based, synthetic biomedical breast cancer datasets, such as Coimbra and Wisconsin, which a publicly available at the (UCI), a machine learning repository.

The BBN was used to identify relevant features strongly connected and dependent on the target variable "diagnosis", achieving 95% coincidence.

The proposed model compared to other data-driven classifier models, i.e., Bernoulli NB, Gaussian NB, AdaBoost, and Ensemble with the proposed model, Bernoulli NB and Gaussian NB. Our study showed that the proposed modeling approach outperformed other models, achieving 98% average accuracy. In addition, the proponents integrated the proposed model into a web-based application prototype, which can be utilized as a support for making decisions for the medical team as a preliminary diagnosis for breast cancer.

Furthermore, in future work, the following the proponents plan for this study: First, consider other datasets to evaluate the proposed model further, significantly increasing the AUC and ROC to 0.9 or 0.1, and lastly, undertake and integrate privacy preservation for the data, as integration of web-based applications in clinical settings necessitates a solid commitment to data privacy. Web-based applications handling clinical data must ensure patient information remains private and secure. However, the study of Rasool *et al*. (2022); Alfian *et al*. (2022) also need their mobile and web application implementations.

Acknowledgment

The authors genuinely appreciate the support and contribution of one another, the main and co-authors, for their motivation, continuous development, and enhancement to finalize valuable input and results of this study.

Funding Information

The authors would like to share their heartfelt gratitude to West Visayas state University-Iloilo.

Author's Contributions

Ma Beth Solas Concepcion: Contributed to the structured introduction, data analysis, and conclusion sections.

Bobby Dioquino Gerardo: Organized idea for methodology; contributed to simulation on stratified stochastic gradient descent algorithm and transcribed analysis for results.

Frank Ibañez Elijorde: Provided data analysis and acquisition; contributed to simulation on Bayesian belief network and transcribed analysis for results.

Joel Traifalgar De Castro: Contributed to simulation on data cleaning and preprocessing.

Nerilou Bermudez Dela Cruz: Conducted related literature review, web app development, and manuscript templating.

We solemnly declare that each author has had the ability and opportunity to make substantive contributions to this research work and we mutually acknowledge and respect the contributions of each author. Collaboration and teamwork have been critical factors in our success in this research endeavor.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and that no ethical issues are involved.

References

Aalaei, S., Shahraki, H., Rowhanimanesh, A., & Eslami, S. (2016). Feature selection using genetic algorithm for breast cancer diagnosis: Experiment on three different datasets. *Iranian Journal of Basic Medical Sciences*, *19*(5), 476.

Aamir, S., Rahim, A., Aamir, Z., Abbasi, S. F., Khan, M. S., Alhaisoni, M., ... & Ahmad, J. (2022). Predicting breast cancer leveraging supervised machine learning techniques. *Computational and Mathematical Methods in Medicine*, *2022*. https://doi.org/10.1155/2022/5869529

Akben, S. B. (2019). Determination of the blood, hormone and obesity value ranges that indicate the breast cancer, using data mining based expert system. *Irbm*, *40*(6), 355-360. https://doi.org/10.1016/j.irbm.2019.05.007

Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., ... & Rhee, J. (2022). Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. *Computers*, *11*(9), 136. https://doi.org/10.3390/computers11090136

CDC. (2022). If you have mammography report say you have dense breasts. *Centers for Disease Control and Prevention*. https://www.cdc.gov/cancer/breast/

Curtin, S. C., & Garnett, M. F. (2023). Suicide and Homicide Death Rates Among Youth and Young Adults Aged 10-24: United States, 2001-2021. https://doi.org/10.15620/cdc:128423

Dag, A. Z., Akcam, Z., Kibis, E., Simsek, S., & Delen, D. (2022). A probabilistic data analytics methodology based on Bayesian Belief network for predicting and understanding breast cancer survival. *Knowledge-Based Systems*, *242*, 108407. https://doi.org/10.1016/j.knosys.2022.108407

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299-310. https://doi.org/10.1109/TKDE.2005.50

Kharya, S., Agrawal, S., & Soni, S. (2014). Using Bayesian belief networks for prognosis and diagnosis of breast cancer. *IJARCCE*, *3*, 5423-7. https://doi.org/10.1155/2022/3813705

Kharya, S., Onyema, E. M., Zafar, A., Wajid, M. A., Afriyie, R. K., Swarnkar, T., & Soni, S. (2022). Weighted Bayesian belief network: A computational intelligence approach for predictive modeling in clinical datasets. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/3813705

Koscielny, M. T. S. (1999). The rationale for early diagnosis of cancer: The example of breast cancer. *Acta Oncologica*, *38*(3), 295-303. https://doi.org/10.1080/028418699431357

Lundervold, A. J., Vik, A., & Lundervold, A. (2019). Lateral ventricle volume trajectories predict response inhibition in older age A longitudinal brain imaging and machine learning approach. *Plos One*, *14*(4), e0207967. https://doi.org/10.1371/journal.pone.0207967

Nanglia, S., Ahmad, M., Khan, F. A., & Jhanjhi, N. Z. (2022). An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. *Biomedical Signal Processing and Control*, *72*, 103279. https://doi.org/10.1016/j.bspc.2021.103279

Naqvi, S. A. A., Tennankore, K., Vinson, A., Roy, P. C., & Abidi, S. S. R. (2021). Predicting kidney graft survival using machine learning methods: Prediction model development and feature significance analysis study. *Journal of Medical Internet Research*, *23*(8), e26843. https://doi.org/10.2196/26843

Ong, H. C. (2011). Improving classification in Bayesian networks using structural learning. *International Journal of Mathematical and Computational Sciences*, *5*(3), 403-407. https://doi.org/10.5281/zenodo.1335262

Peng, Y., & Ding, Z. (2012). Modifying Bayesian networks by probability constraints. *arxiv Preprint arXiv:1207.1356*. https://doi.org/10.48550/arXiv.1207.1356

Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M. R., Qu, Q., & Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. *International Journal of Environmental Research and Public Health*, *19*(6), 3211. https://doi.org/10.3390/ijerph19063211

Sivakami, K., & Saraswathi, N. (2015). Mining big data: Breast cancer prediction using DT-SVM hybrid model. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, *1*(5), 418-429. https://ijseas.com/volume1/v1i5/ijseas20150547.pdf

Wardle, J., Robb, K., Vernon, S., & Waller, J. (2015). Screening for prevention and early diagnosis of cancer. *American Psychologist*, *70*(2), 119. https://doi.org/10.1037/a0037357

Wolberg, W., Street, W., & Mangasarian, O. (1995). Breast cancer Wisconsin (diagnostic) UCI machine learning repository. *Irvine, CA, USA*. ttps://doi.org/10.24432/C5DW2B

Yu, C. S., Lin, Y. J., Lin, C. H., Lin, S. Y., Wu, J. L., & Chang, S. S. (2020). Development of an online health care assessment for preventive medicine: A machine learning approach. *Journal of Medical Internet Research*, *22*(6), e18585. https://doi.org/10.2196/18585