Received: 10 May 2023 | Revised: 12 September 2023 | Accepted: 19 October 2023
DOI: 10.1111/exsy.13498

ORIGINAL ARTICLE

Expert Systems WILEY

Semantic segmentation and classification of polycystic ovarian disease using attention UNet, Pyspark, and ensemble learning model

Ashwini Kodipalli[1,2] | Susheela Devi[1] | Santosh Dasar[3]

[1]Department of Computer Science and Automation, Indian Institute of Science, Bangalore, Karnataka, India

[2]Department of Artificial Intelligence & Data Science, Global Academy of Technology, Bangalore, Karnataka, India

[3]Department of Radiology, SDM College of Medical Sciences & Hospital, Shri Dharmasthala Manjunatheshwara University, Dharwad, India

Correspondence
Ashwini Kodipalli, Department of Computer Science and Automation, Indian Institute of Science, Bangalore, Karnataka, India.
Email: kashwini@iisc.ac.in; dr.ashwini.k@gat.ac.in

Abstract

Ovarian abnormality like polycystic ovarian disease (PCOD) is one of the most common diseases among women worldwide. PCOD not only has an impact on infertility but also hurts the psychological well-being of women affecting their quality of life. In this study, a two-class pattern learning problem is designed for the classification of PCOD. In total, 37 clinical parameters and abdominal ultrasound images of women are collected under the proper ethical protocol. Using only clinical data, an accuracy of 93.7% is obtained using Random Forest as the classifier which is further improved to 95.54% by using a Randomized Search CV during Random Forest classification. The ultrasound images are classified using the proposed Attention-UNet architecture and a mean Dice score of 0.945 is obtained indicating more accurate segmentation. The segmented images are passed through the state-of-the-art EfficientNet B6 for the classification of PCOS and non-PCOS and recorded an accuracy of 95.47%. Using big data architecture Pyspark, the performance is further enhanced to 96.8% and 96.3% for clinical and ultrasound images respectively along with the reduced computational speed. The results of these classifiers are then used to create metadata and a customized Artificial Neural Network is applied for the final prediction of PCOD and non-PCOD. From the results, it can be seen that the stacking model outperformed with an accuracy of 98.12% when compared to the single classifier. Our proposed method has very good performance with less computation, contributing a new architecture to evaluate PCOD and hence helping to improve the wellness of women.

KEYWORDS
attention-UNet, big data, ensemble learning model, metadata, parallel processing, PCOD, Pyspark, stacking

1 | INTRODUCTION

Ovarian carcinoma (OC) is most common in women after breast carcinoma. It is the 5th most common cancer resulting in death. Ovarian tumours are very difficult to detect until they spread to the pelvic and stomach region (Bharati et al., 2020). It is a type of cancer that is asymptomatic in nature and this affects the peritoneum region of ovaries and fallopian tubes. In 2022 reports 295,414 women were subjected to ovarian cancer in the world out of which only 60% of women survived and recovered by various diagnosis methods (Tiwari et al., 2022). Metastasis of tumour cells leads to a high risk to women's life. Based on origin ovarian carcinoma is classified as epithelial and tube ovarian cancer, however, epithelial ovarian cancer is the most common carcinoma accounting for 90% (Bharati et al., 2021). The tumours can be benign or malignant. OC has the

© 2023 John Wiley & Sons Ltd.

highest mortality rate when compared to other gynaecological cancers. By growing technique detection is made easy, though it is asymptomatic. Images can be taken through Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Ultrasound sonography (Danaei Mehr & Polat, 2022). MRI is the best technique to classify malignant and benign tumours in the ovary (Bharati et al., 2021). Using ML and DL algorithms, various models were proposed to detect the tumours in early stages. Some models help in the identification of benign and malignant tumours, and various models focus on diagnosis and treatment measures. If ovarian tumours are detected in the early stage, then recovery is more likely (Bharati et al., 2021). Basic treatment involves chemotherapy and surgery in the long run.

1.1 | Big data tools used for biomedical applications

The volume of information flowing in continuously from different sources is increasing (Thara & Divya, 2021), especially with the development of technology. Big data platforms could be useful in storing and processing the growing volume of data in this situation. A big data platform is a big data management integrated computing solution that combines several software platforms, applications, and hardware. It is a one-stop architecture that handles all of a business's data needs, regardless of the quantity and size of the data at hand. Because of their effectiveness in managing data, enterprises are increasingly using big data platforms to collect a ton of data and transform it into organized, usable business insights (Silva et al., 2022).

Big data platforms that are both open source and commercially available are currently inundating the market. They offer various capabilities and characteristics for application in a big data context. The following crucial characteristics should be present in any good big data platform:

1. Capability of incorporating new software and technologies by changing business requirements.
2. Support for many data formats.
3. Capability to handle enormous amounts of streaming or at-rest data.
4. Possession of a wide range of conversion tools to convert data to various desired formats.
5. The ability to support linear scalability.
6. The capacity to handle data at any speed.
7. The availability of tools for searching through enormous data sets.
8. The capacity for rapid deployment.
9. The availability of tools for data analysis and reporting needs.

The magnitude and complexity of the data, the needs for processing and analysis, and, of course, the budget all play a role in selecting the best big data platform.

The two main objectives of this work are: To detect and classify the presence of polycystic ovarian disease (PCOD) using the clinical parameters and to segment and classify the PCOD using ultrasound images. To achieve these objectives, the following contributions were made:

1. Using the customized attention UNet model, semantic segmentation is carried out on ultrasound images to detect the PCOD.
2. The segmented results obtained from the attention UNet model are further classified using the state-of-the-art convolution neural networks (CNN) models. The results with and without the attention layer are compared.
3. On the clinical parameters, various ensemble techniques are used for the diagnosis of PCOD. The results of the best classifier are further fine-tuned using hyperparameter tuning techniques.
4. One of the main contributions of this paper is the combining of the clinical parameters and the ultrasound images to get better results. As directly combining them is difficult, the results obtained from the classification results of the two data are combined together to form metadata. Artificial Neural Network is used on this metadata to identify the presence or absence of PCOD.

The paper is organized as follows: Section 2 reports the detailed literature survey, Section 3 provides the architecture framework and the methodology, Section 4 gives the detailed implementation of the results and Section 5 concludes the paper.

2 | LITERATURE SURVEY

As there are very limited works carried out in the detection of ovarian abnormalities using big data tools, the entire literature survey is divided into two sections. Section 2.1 describes the detection of ovarian abnormalities using machine learning algorithms and Section 2.2 describes the diagnosis of abnormalities using deep learning.

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.1 | Ovarian cancer using machine learning algorithms

Table 1 gives the list of works using machine learning algorithms. Zheng Zhang et al. (Bharati et al., 2020), in their research work, compared various Machine Learning algorithms like Logical Regression Classifier (LRC), CNN, and Internet of Medical Things (IoMT) for the prediction of ovarian cancer. The author has considered the dataset of Ultra Sound Images to examine the tumours in pregnant women. The Ultra Sound Images through CNN showed an increase in obstetric perinatal and maternal mobility rates. Nasser Taleb et al. (Tiwari et al., 2022) used several Machine Learning algorithms in their research work such as support vector machine (SVM), and K-Nearest Neighbour (KNN). The author used a dataset having 14 parameters for the prediction of ovarian cancer and obtained an accuracy of 98.1% using SVM and 97.16% using KNN. Adithya et al.

TABLE 1 Existing works using machine learning models.

Year of issue	Author	Dataset	Objective	Method	Result
2020	Zheng Zhang et al.	Ultrasound images	To examine the tumours in pregnant women	Logistic regression classifier (LRC), convolution neural networks (CNN), internet of medical things (IoMT)	Ultrasound images employed through CNN show increased obstetric perinatal and maternal mobility rates
2022	Nasser Taleb et al.	14 parameters	Diagnosis for ovarian cancer	Support vector machine (SVM), K-nearest neighbour (KNN)	The proposed model can detect tumours and this model was verified by SVM and KNN with an accuracy of 98.1% and 97.16%
2021	Adithya et al.	Kaggle dataset	To classify benign tumours and malignant tumours	Random forest	The accuracy of the model was the same as the other DL models and better than traditional ML models
2016	Hemitha Pathak et al.	120 patients	To classify benign tumours and malignant tumours	Grey level co-occurrence algorithm (GLCM), support vector machine (SVM), RELIEF-F	The proposed system was fully automated and features were evaluated using the relief-f algorithm and obtained 92% accuracy
2022	Francesca Arezzo et al.	64 patients	Prediction of progression-free survival in patients	Logistic regression, random forest (RFF), K-nearest neighbour (KNN)	Using RFF best ML model was developed for predicting 12-month PFS with 93.7% accuracy.
2022	Jiaojiao Li et al.	470 patients, 1316 radionics features	Classifying type-1 and type-2 epithelial ovarian cancer	Random forest, logistic regression, Naïve Bayes, K-nearest neighbour (KNN), extreme gradient boosting	The combined model using three ML algorithms gave the best performance of 93.4% and was successfully able to differentiate type-1 and type-2
2022	Amir Sorayaie Azar et al.	Four parameters	Predictors of ovarian cancer	Random forest, decision tree, adaptive boosting, K-nearest neighbour (KNN), extreme gradient boosting, support vector machine (SVM)	Using the RFF method an ML model was developed for predicting cancer with an accuracy of 88.72%
2021	Laboni Akter and Nasrin Akhter	PLCO dataset	To predict cysts in the ovary using ultra-sonography	Random forest, K-nearest neighbour (KNN), extreme gradient boosting	Using the RFF method highest accuracy was obtained that is, 99.50%
2018	Oliver Klein et al.	20 patients	Classifying epithelial ovarian cancer	Convolution neural networks (CNN), support vector machine (SVM)	CNN technique was employed with 85% accuracy and therefore classified different types of epithelial cancer
2020	Youg'ai Li et al.	501 women	Differentiating malignant ovarian tumours	ROC, CHI-SQUARE test, diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC)	Using the ML algorithm, a model was developed to differentiate the tumours in patients and obtained an accuracy of 90.2%

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(Danaei Mehr & Polat, 2022) considered the Kaggle dataset to analyse Benign tumours and Malignant tumours using a machine learning algorithm named Random Forest. The accuracy of the model was found to be the same as the other deep learning models and better than traditional machine learning models. Hemitha Pathak et al. (Thara & Divya, 2021) considered several machine learning algorithms like the Grey level co-occurrence algorithm (GLCM), SVM, and RELIEF-F. The author considered a dataset of 120 patients to analyse Benign tumours and Malignant tumours. The author obtained an accuracy of 92% using the RELIEF-F algorithm. Francesca Arezzo et al. (Silva et al., 2022) have worked on several machine learning algorithms such as Logistic Regression, Random Forest (RFF), and KNN to predict the progressive free survival of patients undergoing ovarian cancer. The authors obtained an accuracy of 93.7% using the RFF model that predicted 12 months of Progression Free Survival.

Jiaojiao Li et al. (Bharati et al., 2021) have worked on different machine learning algorithms like Random Forest, logistic regression, Naïve Bayes, KNN, and Extreme gradient boosting considering a dataset having 470 patients and 1316 radio mics features which classifies an image as Type-1 and Type-2 epithelial ovarian cancer. The authors obtained an accuracy of 93.4% using a combined model of three Machine Learning algorithms that could successfully differentiate Type-1 and Type-2 cancer. Amir Sorayaie Azar et al. (Rachana et al., 2021) considered a dataset having four parameters for the prognostication of ovarian cancer using different Machine Learning algorithms like Random Forest, Decision tree, Adaptive boosting, KNN, Extreme gradient boosting, and SVM. The author obtained an accuracy of 88.72% using the RFF method. Laboni Akhter and Nasrin Akhter (Tanwar et al., 2022) used several machine learning algorithms namely Random Forest, KNN, and Extreme gradient boosting for the prediction of cysts in the ovary, considering the PLCO dataset. The author obtained an accuracy of 99.50% through the RFF method. Oliver Klein et al. (Hassan & Mirza, 2020) has considered a dataset of 20 patients for the categorization of epithelial ovarian cancer using different Machine Learning algorithms like CNN and SVM. The author obtained an accuracy of 85% through the CNN technique. Youg'ai Li et al. (Zhang et al., 2021) have considered various evaluation criteria such as ROC, CHI-SQUARE TEST, diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC) to distinguish Malignant tumour considering a dataset of 501 women. The author obtained an accuracy of 90.2% using these ML models.

2.2 | Ovarian cancer using deep learning algorithms

Table 2 lists the works carried out for ovarian cancer prediction using deep learning methods. Ching-Wei-Wang et al. (Khanna et al., 2023) dealt with four kinds of tissue samples to suggest treatment for ovarian cancer. Their research work considered Kaplan–Meier PFS analysis, and the Cox proportional hazard regression model for proposing the AIM2-DL model which inferred reduced development of tumours after a cycle of suggested treatment with an accuracy of 92%. Giacomo Avesani et al. (Baweja & Kanchana, 2023) in their research work examined 218 patients to identify predictors of cancer such as gene mutation. Using CNN a radio mic model was proposed getting an accuracy of 74%. This model identified BRCA gene mutation in patients and subjected them to OC. Tsukasa Saida et al. (Swamy & Nandini Prasad, 2022) considered 365 patients and proposed a diagnosis method for Ovarian cancer. Based on MRI images of patients, a DL algorithm like CNN was employed to suggest a diagnosis method for curing ovarian tumours with a performance of 89%. Yasuyo Urase et al. (Reka & Elakkiya, 2022) examined 50 cases, sparse samples, and CT images to detect the metastasis of ovarian tumours. Using DL algorithms like Residual Encoder-Decoder Convolution Neural Network (RED-CNN) a model was developed to detect metastasis in patients with an accuracy of 95%.

Guangxing Wang et al. (Hdaib et al., 2022) considered 74 patients to identify malignant tumours in patients. In their research work, they employed DL algorithms like Convolution Neural Networks to propose a model which can detect the presence of malignant tumours in patients. Lei Zhang et al. (Sreejith et al., 2022) dealt with Ultrasound images and features of images of CT and MRI for early detection of ovarian cancer. Their work used Uniform Local Binary Pattern, deep learning network, and Cost-sensitive Random Forest to propose a model which detects the lesions formed and malignant tumours in patients. Chengzhu Wu et al. (Inan et al., 2021) considered 988 image samples for classifying tumours formed in patients. They used DL algorithms like CNN to develop a model to classify tumours in patients which were on par with medical staff. Shuo Wang et al. (Guha et al., 2022) examined 245 patients for predicting the serious formation of tumours in ovaries and fallopian tubes. Using DL analysis like Kaplan–Meier PFS analysis, and Cox proportional hazard regression model one can detect metastasis of tumours at serious stages of carcinoma with an accuracy of 71.2%. He-Li-Xu et al. (Bhardwaj & Tiwari, 2022) examined MRI images for the identification of tumours in patients. The model proposed using a DL algorithm like CNN can identify the tumour cells in patients. Rania M. Ghoneim et al. (Prapty & Shitu, 2020) considered Multi-Modal data for the diagnosis of ovarian tumours. Using DL algorithms like the long short-term memory model and CNN, they propose a model which will give ***more precise treatment for Ovarian tumours. Ashwini et al. (Hegde & Kodipalli, 2022; Kodipalli & Devi, 2021; Kodipalli & Devi, 2023; Kodipalli, Devi, et al., 2022; Kodipalli, Fernandes, et al., 2023; Kodipalli, Guha, et al., 2022; Kodipalli, Gururaj, et al., 2023; Ruchitha et al., 2022), contributed extensively to the detection of PCOS using a questionnaire and found that Fuzzy TOPSIS outperformed the SVM algorithm (Kodipalli & Devi, 2021). Watershed and active contour random walker were used in (Ruchitha et al., 2022) for segmenting the ovarian tumour and it was found that the watershed algorithm outperformed the active contour random walker algorithm. The mental condition of women suffering from ovarian cancer was analysed in Kodipalli & Devi, 2023 and Hegde and Kodipalli (2022)) and it was found that women with ovarian cancer have more mental problems compared to women without ovarian cancer. The novel variant of the CNN

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Existing works using deep learning models.

Year of issue	Author	Dataset	Objective	Method	Result
2022	Ching-Wei-Wang et al.	Four kinds of tissue sample	Treatment for ovarian cancer	Kaplan–Meier PFS analysis, Cox proportional hazard regression model	AIM2-DL model developed using the DL algorithm with an accuracy of 92% can distinguish the low recurrence of tumours after treatment.
2022	Giacomo Avesani et al.	218 patients	Predictors like gene mutation in cancer patients	Convolution neural network (CNN)	Radiomic models helped in predicting BRCA mutation at the early stages of cancer using CNN with an accuracy of 74%
2022	Tsukasa Saida et al.	365 patients	Diagnosis for ovarian tumours	Convolution neural network (CNN)	CNN method suggested some diagnosis methods based on MRI with an accuracy of 89%
2020	Yasuyo Urase et al.	50 test cases, Sparse-sampling and CT images	Detection of metastasis in ovarian tumours	Residual encoder-decoder convolution neural network (RED-CNN)	Using the DL algorithm a model was developed to detect metastasis with an accuracy of 95%.
2022	Guangxing Wang et al.	74 patients	Identification of tumours in patients	Convolution neural network (CNN)	Using CNN 99.7% accuracy is obtained in identifying malignant tumours
2019	Lei Zhang et al.	Ultrasound image features	Early detection of ovarian carcinoma	Uniform local binary pattern, deep learning network, cost-sensitive random forest	Results inferred distinct lesions in malignant and benign tumours in patients
2018	Chengzhu Wu et al.	988 image samples	Classification of tumours based on ultrasound images	Convolution neural network (CNN)	Using DL algorithm classification was on par with medical staff.
2019	Shuo Wang et al.	245 patients	Prediction of serious tumours in the ovary	Kaplan–Meier PFS analysis, Cox proportional hazard regression model	Using DL analysis one can detect metastasis of tumours at the serious stage of carcinoma with an accuracy of 71.2%
2022	He-Li-Xu et al.	MRI images	Identification of tumours in patients	Convolution neural network (CNN)	The model proposed using the DL algorithm can identify the tumour cells
2021	Rania M. Ghoneim et al.	Multi-modal data	Diagnosis for ovarian tumours	Long short-term memory model, convolution neural network (CNN)	The proposed model will give more precise treatment for ovarian tumours

model to classify ovarian tumours as benign or malignant was proposed in (Kodipalli, Guha, et al., 2022) and the proposed model outperformed the state-of-the-art architectures of 2014 winning architectures of ILSVRC. The novel single pipeline architecture for segmenting and classifying the tumours was proposed in (Kodipalli, Devi, et al., 2022) which not only outperformed in terms of time but also accuracy. The novel inverted fuzzy c means architecture for the accurate detection of ovarian tumours was proposed in (Kodipalli, Fernandes, et al., 2023). The performance of UNet and Transformers was compared in (Kodipalli, Gururaj, et al., 2023) and it was observed that Transformers performed better in malignant lesion detection.

2.3 | Motivation

The process of working with large medical datasets which contain multiple dimensions and analysing the patterns in the data is tedious in itself but is crucial. In the current trend, while PCOD is a prevalent disease, it is necessary to dig into the root cause of the disease itself which can be done through extensive data mining and identifying the correlation between attributes. Although similar work has been done using basic machine learning algorithms and deep learning, the effectiveness of analysis increases when the metadata is created and prediction is done using metadata. Big data tools like Pyspark and Hadoop have been used as they not only provide accurate predictions but also trace the factors affecting each individual based on these factors. These big data tools are built to handle such large complex datasets, process, store, and analyse data from various sources and provide insightful instances based on genetic information, lifestyle factors, and relevant features. It can be observed that by the combination of Pyspark, Hadoop, and ML algorithms, the effectiveness of predicting PCOS has increased drastically while also providing early detection patterns, better diagnosis, optimized treatment, and improved patient response. PCOD being a lifestyle disorder cannot be trained by just a few samples and it is crucial to identify underlying patterns to provide the best treatment to a patient and can be done by the collaboration between big data and machine learning.

3 | METHODOLOGY

3.1 | Dataset

The dataset used in the research work is obtained from SDM College of Medical Sciences and Hospital, Dharwad. There were approximately 5000 patients' data collected over the past 5 years. From each patient, both clinical data and ultrasound images were collected. The entire process is approved by the ethical clearance committee of the hospital. The clinical data obtained is in the CVS file containing 4700 rows and 37 columns of data for predicting PCOS and 4700 ultrasound-scanned images for detecting PCOD. CSV file contains the attributes such as regularity in the periods, cycle length, and hormonal test parameters such as TSH, FSH, and so forth for the initial screening of the PCOD and further, the presence of PCOD is confirmed with the processing of ultrasound images.

3.2 | System design

The current work aims to develop the prediction model for the segmentation and classification of PCOD using clinical data and using ultrasound images. This study includes women of age between 15 and 30 who visited SDM College of Medical Sciences and Hospital, Dharwad. Women who were suffering from PCOD and had a history of PCOD in the family who visited the hospital for checkups and follow-ups were included in the study. The ethical approval committee of the hospital has approved the study and the data has been collected from the hospital. This study involved the radiologist annotating the lesion of interest in the ultrasound images and the gynaecologist providing the clinical parameters data. Women who are having serious illnesses related to the ovaries such as ovarian cancer and pelvic cancer are excluded from this study. The implemented results are validated by an expert practitioner.

3.3 | Architecture framework

3.3.1 | Segmentation

The segmentation architecture is shown in Figure 1.

The segmented model used the UNet architecture and integrated the attention model with the decoder units. Further, we proposed an attention model with the decoder blocks to enhance the segmentation prediction. The architecture consists of a skip connection which reduces the inconsistency in feature learning. The model also has spatial attention with sequential channels. The overall proposed architecture is shown in Figure 1. The sequential channel and the spatial attention are used to enhance the clarity of the encoding module by introducing the attention units to generate channel attention focusing on “what” is meaningful given an input image and spatial attention which focuses on “where” is the informative region by exploring inter and intra-feature relationships. Features that give information about the geometry, size, and shape of the follicles hold an important role in the detection of the presence of PCOD. Along with these features, the details like histogram features, centroid coordinates, and Eigenvalues of the image including entropy, and skewness are also extracted for clear segmentation of PCOD.

3.3.2 | Big data

The data is given to the Kafka producer and is processed in parallel using Pyspark and is collected at the consumer end using Kafka consumer. During the data processing, the ensemble learning algorithm is applied to the dataset which is a pre-trained model in the consumer end that predicts PCOD. The ensemble classification process assigns the class to the incoming data and thus using Pyspark, apache Kafka with ML classifier, parallel processing is achieved with low latency, low runtime, and high accuracy.

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

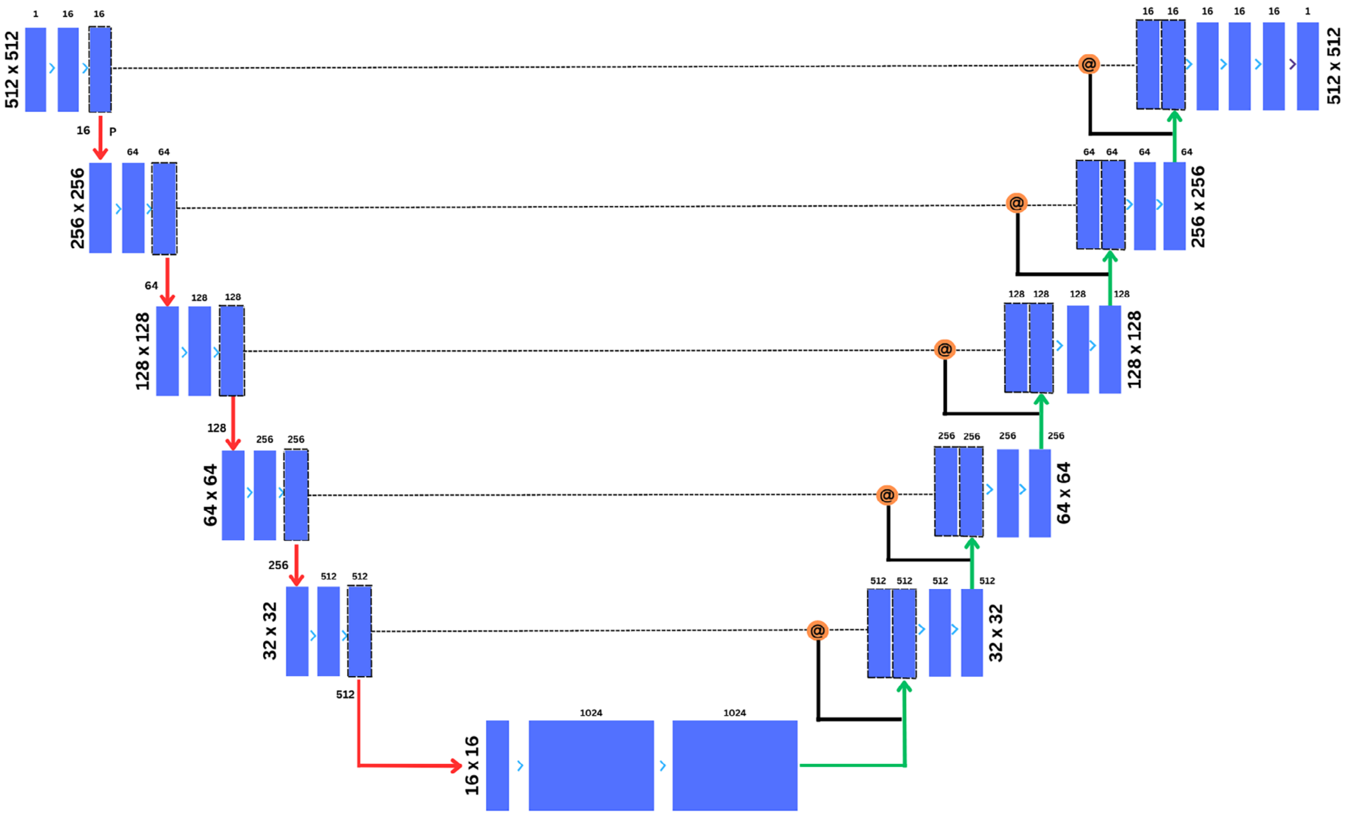

FIGURE 1 Customized U-Net attention model.

Pyspark is thus a powerful tool to predict PCOS as it integrates data processing, machine learning, and parallel processing via map reduction and scalability while interacting with the Python ecosystem. Pyspark offers tools and libraries and provides methods including correlation analysis, principal component analysis (PCA), and feature hashing. Pyspark integrates with MLlib, Spark's inbuilt ML library, which provides a range of distributed machine learning algorithms that can be applied to the preprocessed PCOS datasets to build predictive models. Pyspark then deploys these saved models on large-scale datasets and does parallel processing for faster computation. ML algorithms such as random forests, decision trees, CNN, and support vector machines can be utilized to predict the likelihood of PCOS. This enables rapid PCOS prediction by efficient processing of massive medical datasets, enhanced feature engineering, scalable machine learning model training, and real-time analytics.

The working of Pyspark is as follows. Data is stored in HBase and accessed by the consumer by reading the data through HBase. HBase is a column-based NOSQL data store that enables the data to be accessed randomly by using basic queries. This solves the problem inherent in Hadoop of accessing the data sequentially.

Using this HBase we can perform multiple read-write operations. Data read through HBase is given to the Hadoop streaming infrastructure that reads data, combines them into streams, and processes it using the Map-Reduce pattern. Map reduce is a sequence of two tasks: map and reduce. The map task reads the data and converts them into key-value pairs. Reduce task combines the output thus obtained and reduces the amount of data. These streaming-processing techniques are mainly used to reduce time complexity, reduce latency, enable fast computations, and reduce the amount of data to be processed. By applying partitioning, we can increase the number of reduced operations and by applying combining we can reduce output much more efficiently and quickly. To improve the parallelism, we use pipelines wherein the output of one reduce serves as input to a conjoined map. This process of computation is made to run in multiple nodes to enable complete utilization of resources and enable multi-threaded parallel processing. This processed data is stored in HBase. Producers can access this data directly from HDFS or through HBase. The architecture framework is shown in Figure 2.

Map Reduce when integrated with Pyspark provides features such as fault tolerance and efficient cluster management as it uses HDFS and complex API thus increasing computing capacity. The use of map reduce in our work appears in various steps. Initially, the data is divided and distributed across nodes in a cluster to provide faster computation via scaling by distributing the workload. In each of these nodes, data is cleaned and preprocessed. It undergoes feature extraction, dimensionality reduction, and PCA. For multiple iterations-based algorithms, each iteration gets subsequently carried out in different nodes thus speeding up the training process. Thus, map reduce helps speed up the entire process by enabling parallel distribution and maximum utilization of computational resources.

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

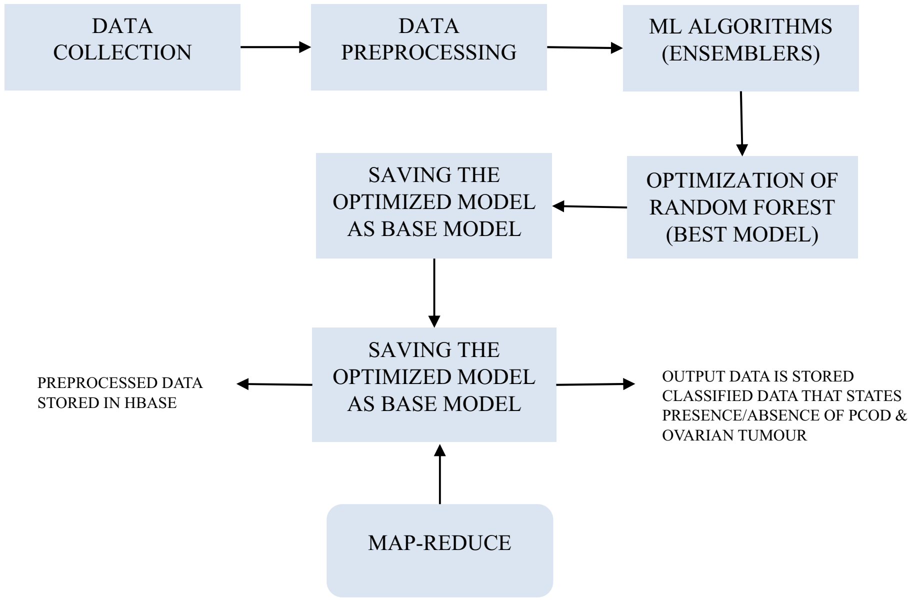
DATA COLLECTION
DATA PREPROCESSING
ML ALGORITHMS (ENSEMBLERS)
OPTIMIZATION OF RANDOM FOREST (BEST MODEL)
SAVING THE OPTIMIZED MODEL AS BASE MODEL
SAVING THE OPTIMIZED MODEL AS BASE MODEL
PREPROCESSED DATA STORED IN HBASE
OUTPUT DATA IS STORED CLASSIFIED DATA THAT STATES PRESENCE/ABSENCE OF PCOD & OVARIAN TUMOUR
MAP-REDUCE

FIGURE 2 Architecture framework for the prediction of polycystic ovarian disease (PCOD) using the Apache Spark.

3.3.3 | Proposed architecture framework using stacking

The current research work uses the metadata for the accurate prediction of PCOD and non-PCOD. For accurate prediction, the predictions from the Random Forest, SVM, EfficientNet B6, and NASNet Large models are combined using the Maximum Voting combination selection scheme. The labels from the original dataset are augmented to the metadata dataset. On the newly generated metadata, the customized ANN model with four hidden layers of varying number of neurons in the hidden layer is used. For better performance, the dropout value of 0.2 is used at the input and at all the hidden layers. The detailed architecture framework is given in Figure 3.

4 | RESULTS AND DISCUSSION

The results are represented in five sections. Section 4.1 provides the implementation details. Section 4.2 describes the segmentation and classification of the ultrasound images using a computational model. Section 4.3 describes the prediction of PCOD using various ensemble learning models. Section 4.4 describes the comparison of the results with and without using Big data tools. Section 4.5 provides the stacking results.

4.1 | Implementation details

Using bicubic interpolation, all the images are resized to 512 × 512 pixels. The batch size was set to 64 and every model was trained up to 1000 epochs. A final fully connected layer with ReLU activation function has 256 hidden neurons followed by a dropout layer with a probability of 0.5 to prevent overfitting. Adam optimizers are used with the parameter values (beta 1 and beta 2) set to 0.6, 0.8, and the learning rate is 0.0001. The last dense layers of all architecture are modified to output two classes corresponding to benign and malignant. All pre-trained CNN models are fine-tuned separately. The training and testing of the proposed architecture is implemented using Python using the Keras package and run on Nvidia RTX 3060 GPU with 32 GB RAM.

4.2 | Segmentation and classification results of ultrasound images for the detection of PCOD

The metrics used to measure the performance of the segmentation are the dice score and the Jaccard score. The dice score is given below:

$$\text{Dice} = \frac{2\times|S\cap G|}{|S|+|G|},$$

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

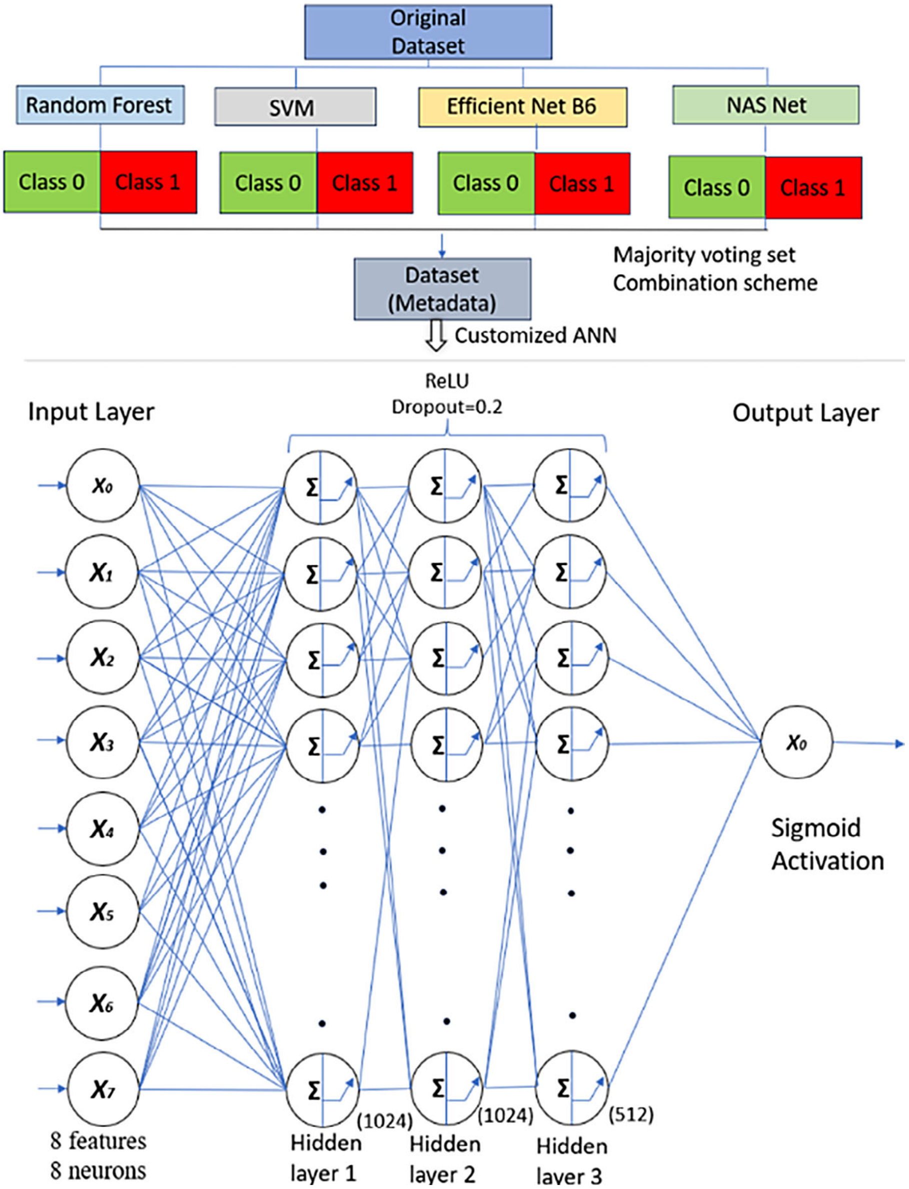

FIGURE 3 Detailed architecture framework.

where *G* is the ground truth and S is the segmented image.

$$\text{Jaccard} = \frac{\text{dice}}{2-\text{dice}}.$$

The evaluated values of Dice and Jaccard scores will be in the range of 0 and 1. Higher the value the better the segmentation result is. Figure 4 provides the segmented results using the Attention-UNet model.

From Table 3, it can be inferred that the proposed attention with the UNet model has outperformed the UNet model without attention. The UNet model performed well for those images where the PCOD has grown well and showed less performance in those images where the PCOD is uncertain. The CNN models—NASNetLarge and EfficientNet B6 are used for the classification of the segmented PCOD. The computed results are shown in Table 4.

The segmented results from the U-Net Attention model are classified using the state-of-the-art CNN models—NASNetLarge and EfficientNet B6. Then the segmented tumours are passed to the classifier, from Table 4, it is observed that the classifier performed well on the Attention-U-Net model and produced the results the accuracy of 94.78% and 95.47% respectively by NASNetLarge and EfficientNet B6.

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

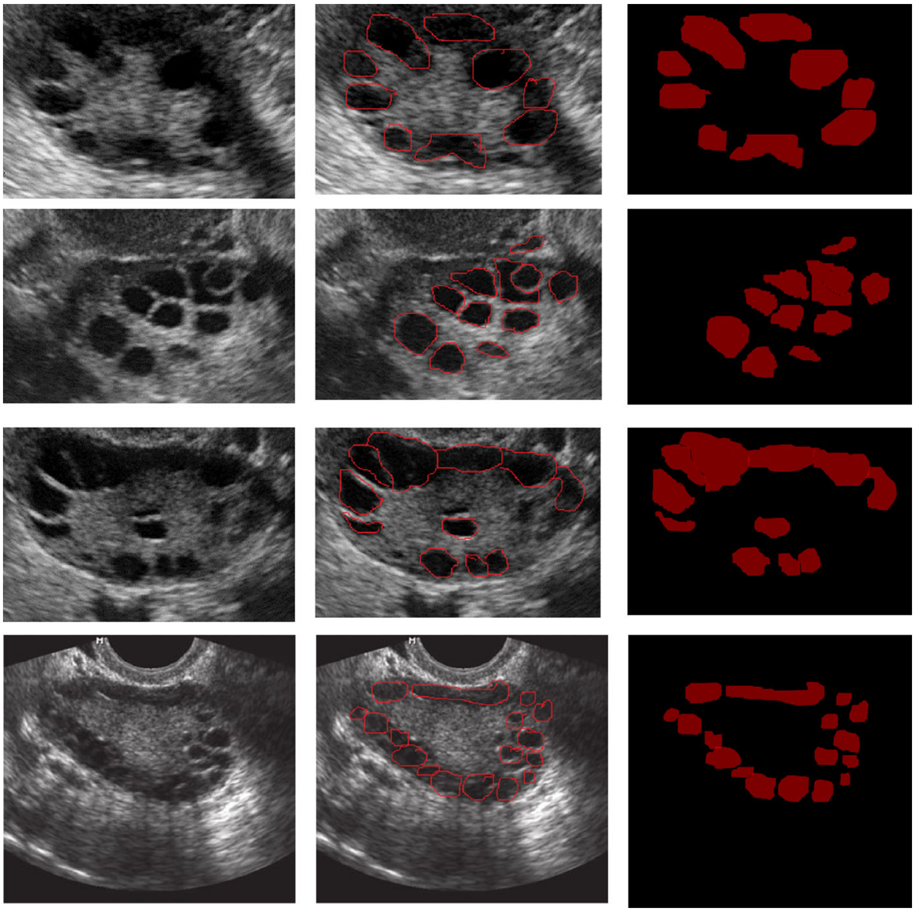

FIGURE 4 The segmented results using the attention-UNet model: input image (Left), ground truth (middle), and segmented results (right).

TABLE 3 Quantitative comparison of average dice and Jaccard score of the UNet model with and without attention gates.

	Attention—UNet				UNet			
	Dice		Jaccard		Dice		Jaccard	
	Background	Foreground	Background	Foreground	Background	Foreground	Background	Foreground
Mean	0.9966	0.945	0.9933	0.9107	0.9875	0.939	0.9833	0.9098
Median	0.99664	0.99331	0.99331	0.96397	0.98773	0.98779	0.98559	0.95784
Std. Dev	0.00091	0.1039	0.00181	0.14989	0.00089	0.1055	0.00176	0.14122

TABLE 4 Classification results of different state-of-the-art convolution neural networks classifiers.

	Attention—UNet			UNet		
Algorithms	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
NASNetLarge	94.78	95.45	95.38	93.26	92.57	72.37
EfficientNet B6	95.47	95.81	94.10	94.13	94.00	89.78

4.3 | Prediction of PCOD using machine learning algorithms

The metrics used to measure the performance of the model are Accuracy, Precision, Recall, and *F*1 score. Mathematically the metrics are described below:

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100,$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100,$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100,$$

$$F1\,\text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}.$$

The performance of the various machine-learning algorithms is shown in Table 5 and Figure 5.

From Table 5 and Figure 5, it is observed that Random Forest has performed the best when compared with ensemble learning models. Further, the results are improved with the hyperparameter tuning techniques. The different tuning techniques used are randomized search CV, Grid search CV, and Bayesian technique. Table 6 describes the improved results using tuning techniques. From the table, it is observed that the results are improved and the Bayesian tuning technique has increased the performance of the Random Forest.

TABLE 5 Performance of various machine learning ensemble model.

Sl. No	Learning model	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
1.	Random Forest	93.7	93.5	92.67	92.12
2.	GradientBoost	89.54	89.88	89.91	89.59
3.	AdaBoost	84.37	84.54	84.67	84.19
4.	XGBoost	88.69	88.49	88.63	88.55

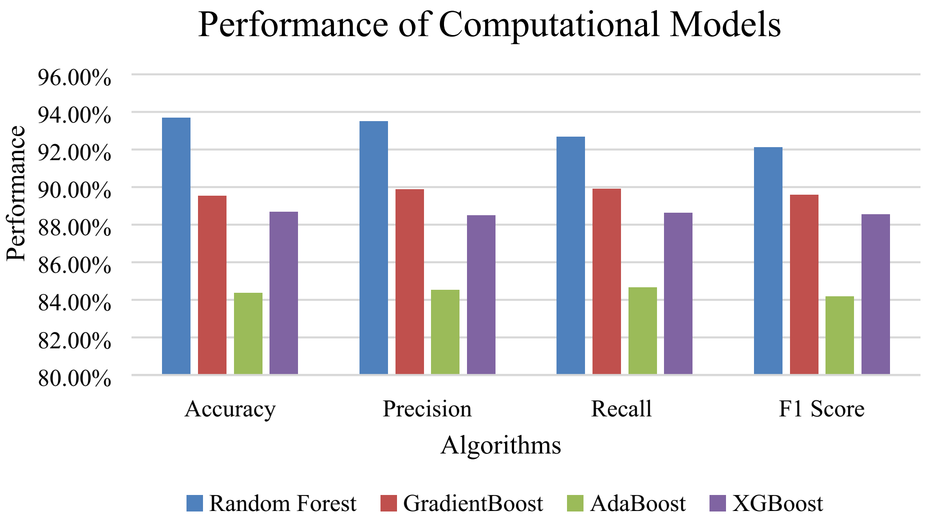

FIGURE 5 Performance of computational models.

TABLE 6 Improved Random Forest results using optimization techniques.

		Random Forest			
Sl. no.	Optimization techniques	Accuracy (%)	Precision (%)	Recall (%)	F1 score
1.	Randomized search CV	95.54	95.67	95.23	95.49
2.	Grid search CV	94.62	95.89	95	94.2
3.	Bayesian	96.57	96.89	96.12	96.28

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.4 | Comparison of the results on the big data platform

The ML model running on the big data platform provided 97.8% of accuracy in predicting the PCOD using clinical parameters; and 96.3% of accuracy in detecting the PCOD using ultrasound images. Because of Pyspark, batch processing is achieved and high accuracy is obtained in less than a few minutes. The results of big data are then compared without using a big data platform. From Table 7, it is observed that the results not only underperformed yet higher time consumption in prediction with the result of 95.54% accuracy in predicting the PCOD using clinical parameters and 95.47% accuracy in detecting the PCOD using Ultrasound images.

The collaboration between Spark and ML has caused significant changes in the results obtained. The main reason for this noticeable increase is that PCOS in most women relies on multiple medical, nonmedical/lifestyle factors and not just two or three factors, thus making it more important to have a deep understanding of the root cause of PCOS to facilitate the patient to get the right treatment based on their issues. It becomes crucial to identify the hidden patterns and recognize anomalies in the data which is highly effective when big data tools are used with ML. By employing Hadoop, Spark, and Map Reduce, we discover valuable information that would otherwise go untraced using basic ML due to the reduced computational demands of the complex dataset by the use of big data (BD) tools. These BD tools use statistical modelling methods to improve the accuracy and distribute the load thus ensuring better performance without overlooking the accuracy of prediction.

4.5 | Stacking model

The Majority Voting combination scheme is used to create metadata from the predictions of models such as Random Forest, SVM, EfficientNet B6, and NASNet large. The dataset contains eight features. On this metadata, the customized artificial neural network which is a meta learner was applied for the prediction of PCOS and non-PCOS. From the results, it is inferred that the customized ANN model with four hidden layers and with a dropout of 0.2 has outperformed the single classifiers as shown in Table 8.

5 | CONCLUSION

In this research work, we analyse the different methods to determine the presence of PCOD and ovarian tumours in women. As we know early diagnosis and detection can help in prevention and cure. This work is purely based on detecting PCOD using machine learning algorithms in integration with big data tools.

First, we apply different ML algorithms to our dataset to check the speed and accuracy of computation, then optimize it using Randomized Search, Grid Search, and Bayesian methods. As we have observed from the results obtained, Random Forest provides better accuracy after optimization. This is the base model that is being used to process the data in Hadoop.

Hadoop supports many data types, that includes structured, semi-structured, and unstructured data, thus making it easy to analyse a large variety of data like text or sensor photos, or medical images. As Hadoop interfaces with many other big data technologies, it enables more sophisticated data processing like machine learning and real-time processing of data. Thus, we can extend this research to real-time data for instant diagnosis based on the model and data saved during the training phase. In Hadoop data is stored in HBase. Data is accessed by consumers and the streaming process starts. Data is made to process in parallel by using Map Reduce and the base model which is saved is applied to it. Data is read, processed, classified, and stored in HBase. HBase provides easy random access to users. Producers can read the output from HDFS or HBase. Using this method of data processing using Hadoop, data is processed quickly, efficiently, with low latency, and with high accuracy. The accuracy

TABLE 7 Comparison of the results with big data and without big data.

Objective	Model	Using apache spark as a platform (%)	Without using apache spark as a platform (%)
Predicting the PCOD using clinical parameters	Random Forest	96.8	95.54
Predicting the PCOD using ultrasound images	EfficientNet B6	96.3	95.47

TABLE 8 Performance of the stacking model.

Model	Accuracy (%)	Precision (%)	Recall (%)	*F*1 score (%)
Stacking	98.12	97.88	97.16	96.89

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of prediction increases significantly and a very large amount of data can be computed due to its distributed computing architecture. Further, the predictions from all the classifiers are combined using a majority voting combination scheme, and the metadata is created. The predictions from the metadata provided an accuracy of 98.12% and outperformed all the single classifiers. This approach gave a new direction in improving the accuracy of the prediction model.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Ashwini Kodipalli https://orcid.org/0000-0001-6549-1056

REFERENCES

Baweja, A. K., & Kanchana, M. (2023). Prediction of polycystic ovarian syndrome using machine learning techniques. In *Machine learning, image processing, network security and data sciences: Select proceedings of 3rd international conference on MIND 2021* (pp. 53–63). Springer Nature Singapore.

Bharati, S., Podder, P., & Mondal, M. R. H. (2020). Diagnosis of polycystic ovary syndrome using machine learning algorithms. In *IEEE region 10 symposium (TENSYMP)* (pp. 1486–1489). IEEE.

Bharati, S., Podder, P., Mondal, M. R. H., Surya Prasath, V. B., & Gandhi, N. (2021). Ensemble learning for data-driven diagnosis of polycystic ovary syndrome. In *International conference on intelligent systems design and applications* (pp. 1250–1259). Springer International Publishing.

Bhardwaj, P., & Tiwari, P. (2022). Manoeuvre of machine learning algorithms in healthcare sector with application to polycystic ovarian syndrome diagnosis. In *Proceedings of academia-industry consortium for data science: AICDS 2020* (pp. 71–84). Springer Nature Singapore.

Danaei Mehr, H., & Polat, H. (2022). Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. *Health and Technology*, *12*(1), 137–150.

Guha, S., Kodipalli, A., & Rao, T. (2022). Computational deep learning models for detection of COVID-19 using chest X-ray images. In *Emerging research in computing, information, communication and applications: Proceedings of ERCICA 2022* (pp. 291–306). Springer Nature.

Hassan, M. M., & Mirza, T. (2020). Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome. *International Journal of Computers and Applications*, *975*, 8887.

Hdaib, D., Almajali, N., Alquran, H., Mustafa, W. A., Al-Azzawi, W., & Alkhayyat, A. (2022). Detection of polycystic ovary syndrome (PCOS) using machine learning algorithms. In *5th international conference on engineering technology and its applications (IICETA)* (pp. 532–536). IEEE.

Hegde, H. S., & Kodipalli, A. (2022). Machine learning based approach for breast cancer detection. In *International conference on computing, communication, and intelligent systems (ICCCIS)* (pp. 782–786). IEEE.

Inan, M. S. K., Ulfath, R. E., Alam, F. I., Bappee, F. K., & Hasan, R. (2021). Improved sampling and feature selection to support extreme gradient boosting for PCOS diagnosis. In *IEEE 11th annual computing and communication workshop and conference (CCWC)* (pp. 1046–1050). IEEE.

Khanna, V. V., Chadaga, K., Sampathila, N., Prabhu, S., Bhandage, V., & Hegde, G. K. (2023). A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. *Applied System Innovation*, *6*(2), 32.

Kodipalli, A., & Devi, S. (2021). Prediction of PCOS and mental health using fuzzy inference and SVM. *Frontiers in Public Health*, *9*, 789569.

Kodipalli, A., & Devi, S. (2023). Analysis of fuzzy-based intelligent health care application system for the diagnosis of mental health in women with ovarian cancer using computational models. *Intelligent Decision Technologies*, *17*, 1–12.

Kodipalli, A., Devi, S., Dasar, S., & Ismail, T. (2022). Segmentation and classification of ovarian cancer based on conditional adversarial image to image translation approach. *Expert Systems*, *10*, e13193.

Kodipalli, A., Fernandes, S. L., Dasar, S. K., & Ismail, T. (2023). Computational framework of inverted fuzzy C-means and quantum convolutional neural network towards accurate detection of ovarian tumors. *International Journal of E-Health and Medical Communications*, *14*(1), 1–16.

Kodipalli, A., Guha, S., Dasar, S., & Ismail, T. (2022). An inception-ResNet deep learning approach to classify tumours in the ovary as benign and malignant. *Expert Systems*, *10*, e13215.

Kodipalli, A., Fernandes, S. L., Gururaj, V., Shriya, V. R., & Dasar, S. (2023). Performance analysis of segmentation and classification of CT scanned ovarian tumours using U-Net and deep convolutional neural networks. *Diagnostics*, *13*, 2282.

Prapty, A. S., & Shitu, T. T. (2020). An efficient decision tree establishment and performance analysis with different machine learning approaches on polycystic ovary syndrome. In *23rd international conference on computer and information technology (ICCIT)* (pp. 1–5). IEEE.

Rachana, B., Priyanka, T., Sahana, K. N., Supritha, T. R., Parameshachari, B. D., & Sunitha, R. (2021). Detection of polycystic ovarian syndrome using follicle recognition technique. *Global Transitions Proceedings*, *2*(2), 304–308.

Reka, S., & Elakkiya, R. (2022). Early diagnosis of poly cystic ovary syndrome (PCOS) in young women: A machine learning approach. In *IEEE international symposium on mixed and augmented reality adjunct (ISMAR-adjunct)* (pp. 286–288). IEEE.

Ruchitha, P. J., Sai, R. Y., Kodipalli, A., Martis, R. J., Dasar, S., & Ismail, T. (2022). Comparative analysis of active contour random walker and watershed algorithms in segmentation of ovarian cancer. In *International conference on distributed computing, VLSI, electrical circuits and robotics (DISCOVER)* (pp. 234–238). IEEE.

Silva, I. S., Ferreira, C. N., Costa, L. B. X., Sóter, M. O., Carvalho, L. M. L., Albuquerque, J. d. C., Sales, M. F., Candido, A. L., Reis, F. M., Veloso, A. A., & Gomes, K. B. (2022). Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. *Journal of Endocrinological Investigation*, *3*, 1–9.

Sreejith, S., Nehemiah, H. K., & Kannan, A. (2022). A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier. *Healthcare Analytics*, *2*, 100102.

Swamy, S. R., & Nandini Prasad, K. S. (2022). Hybrid machine learning model for early discovery and prediction of polycystic ovary syndrome. In *Second international conference on advanced technologies in intelligent control, environment, computing & communication engineering (ICATIECE)* (pp. 1–8). IEEE.

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Tanwar, A., Jain, A., & Chauhan, A. (2022). Accessible polycystic ovarian syndrome diagnosis using machine learning. In *3rd international conference for emerging technology (INCET)* (pp. 1–6). IEEE.

Thara, L., & Divya, T. M. (2021). Detection and prediction system for polycystic ovary syndrome using structural normalized square similarity detection approach. *Natural Volatiles & Essential Oils Journal*, *8*, 2834–2842.

Tiwari, S., Kane, L., Koundal, D., Jain, A., Alhudhaif, A., Polat, K., Zaguia, A., Alenezi, F., & Althubiti, S. A. (2022). SPOSDS: A smart polycystic ovary syndrome diagnostic system using machine learning. *Expert Systems with Applications*, *203*, 117592.

Zhang, X., Liang, B., Zhang, J., Hao, X., Xu, X., Chang, H. M., Leung, P. C. K., & Tan, J. (2021). Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening. *Molecular and Cellular Endocrinology*, *523*, 111139.

AUTHOR BIOGRAPHIES

Ashwini Kodipalli, currently working as an Associate Professor and Heading the Department of AI & Data Science, at the Global Academy of Technology. Currently pursuing Ph.D. from Indian Institute of Science, Bangalore. Her areas of research include Biomedical Image Analysis using Advanced Deep Learning models, and the psychological well-being of women with PCOS and ovarian cancer working in association with NIMHANS and SDM Hospital, Dharwad. She is actively involved in the research by publishing more than 70+ papers in the reputed IEEE and Springer Lecture Notes series conferences and 35+ reputed International Journals. Apart from research activities, she is actively involved in conducting various value-added courses for enriching student skills in the areas of Machine Learning and Deep Learning. She guides a number of students at master's, and UG levels.

Susheela Devi, is a Principal Research Scientist at the Department of Computer Science and Automation at the Indian Institute of Science (IISc) Bangalore. She works in the areas of data mining, machine learning, artificial intelligence, and soft computing. She teaches the courses data mining, data structures and algorithms, artificial intelligence, intelligent agents, computational methods of optimization, soft computing, etc. She guides a number of students at the master's and Ph.D. levels. She has a number of papers in international journals and conferences. She has attended, been a program committee member, and given keynote addresses at a number of international conferences. She has also reviewed papers for a number of reputed international conferences and journals. She has also evaluated books and Ph.D. theses. She has written three books in the field of pattern recognition and machine learning and prepared a web course from NPTEL. To help the community and ML enthusiasts, she has also given a number of talks at various college events and technical conferences.

Santosh Dasar is working as Professor and Radiologist in SDM College of Medical Sciences and Hospitals. His areas of research include analysis of adnexal masses using ultrasound, MRI and CT images Radiology Nuclear Medicine and Medical Imaging.

How to cite this article: Kodipalli, A., Devi, S., & Dasar, S. (2024). Semantic segmentation and classification of polycystic ovarian disease using attention UNet, Pyspark, and ensemble learning model. *Expert Systems*, *41*(3), e13498. https://doi.org/10.1111/exsy.13498

14680394, 2024, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13498 by The Librarian, Wiley Online Library on [20/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License