Comparison of Deep Learning and Traditional Machine Learning Techniques for Classification of Pap Smear Images

1st Abdurrahim Yilmaz
Yildiz Technical University
Mechatronics Engineering
Istanbul, Turkey
a.rahim.yilmaz@gmail.com

2nd Ali Anil Demircali
Yildiz Technical University
Mechatronics Engineering
Istanbul, Turkey
aanil@yildiz.edu.tr

3rd Sena Kocaman
Robert College
Istanbul, Turkey
senakocaman@gmail.com

4th Huseyin Uvet
Yildiz Technical University
Mechatronics Engineering
Istanbul, Turkey
huvet@yildiz.edu.tr

Abstract—A comprehensive study on machine and deep learning techniques for classification of normal and abnormal cervical cells by using pap smear images from Herlev dataset results are presented. This dataset includes 917 images and 7 different classes. All techniques used in this study are modeled by using Google Colab platform with scikit-learn and Keras library inside TensorFlow. In the first study, traditional machine learning methods such as logistic regression, k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Decision Tree, Random Forest and eXtreme Gradient Boosting (XGBoost) are used and compared with each other to find binary classification as normal and abnormal cervical cells. Better results are observed by XGBoost and kNN classifiers among the others with an accuracy of 85%. In the second study, a deep learning model based on Convolutional Neural Network (CNN) is used for the same dataset. Accordingly, accuracies of 99% and 93% are obtained for the training and the test dataset, respectively. CNN model extracts its features from raw data without any label or a feature dataset. In this model, it takes 50 epochs to have these accuracies within 20 minutes of computational time.

Index Terms—Pap smear image, convolutional neural network, machine learning, cervical cancer

I. INTRODUCTION

Cervical cancer is one of the most influential cancer type for women worldwide according to Global Cancer Observatory. It is reported that 570,000 new cases and 311,000 deaths are occurred in 2018 due to this cancer [1]. It is important to find a related diagnosis and find a cure; therefore, machine and deep learning approaches can be used to detect various disease types such as cervical cancer, onychomycosis, breast cancer and etc. by using robust models [2], [3]. To do this, a dataset must be generated for automated illness classification. For automated classification of cervical cancer, Herlev dataset is created from pap smear images and each image of this dataset has been labeled initially. Thus, it can be used for both machine and deep learning techniques since machine learning techniques require labeled images for training. On the other hand, in this dataset, seven cervical cell types are available such as superficial squamous epithelial, intermediate squamous epithelial, columnar epithelial, mild dysplasia, moderate dysplasia, severe dysplasia, and carcinoma-in-situ. These types must be converted from seven types to binary classes as normal and abnormal cells for the classification of seven classes. In this dataset, abnormal cervical cells are mild dysplasia, moderate dysplasia, severe dysplasia, and carcinoma-in-situ. In the meantime, normal cervical cells are superficial squamous epithelial, intermediate squamous epithelial, columnar epithelial [4].

Jantzen et al. [4] studied the classification of Herlev dataset with linear least-squares classifier with an accuracy of 94%. However, this accuracy stands lower for current methods that have better metrics by using pretrained CNN especially deep learning algorithms [5]. Marinakis et al. [6] classified cervical cells by using genetic algorithm for segmentation and meta-heuristic algorithm and nearest neighbor based classification with 20 numerical features. Indicated results with an accuracy of 89% showed that this accuracy is not adequate compared with the current machine learning techniques such as the fuzzy C-means method [7]. Chankong et al. [7] developed a novel method to segmentate for pap smear images with an accuracy of 99%. This accuracy is achieved by using patch-based fuzzy C-means (FCM) and FCM clustering method for segmentation and FCM algorithm for classification with features that are obtained from cytoplasm, nucleus and background. Although this is the highest accuracy for binary classification of pap smear images by using the same dataset, the data preprocessing and the segmentation process take a long time and are challenging processes. Ashok et al. [8] used multi-thresholding method for segmentation and SVM classifier with 99% accuracy for a private dataset with 14 texture and 30 shape features. The features are obtained for a different dataset that includes 150 images. Though, the accuracy level is higher, no information is available regarding to complicated and larger dataset. Zhang et al. [5] proposed a deep convolutional neural networks for classification of pap smear images with (256 × 256 pixels) image size. The proposed method uses transfer learning by using ImageNet dataset to set the weights of deep neural network efficiently. To the best of our knowledge, this model has 98% accuracy with the highest accuracy for the studies related with pap smear classification. In deep learning, feature
extraction is an automated process unlike machine learning. This property reduces the model developing process. However, average training time progresses slowly, which takes 4 hours for 30 epochs and 256 × 256 pixels image size in this study for each training effort.

The significance of this paper is to show how the models are tested and improved through optimization and reduced computing time using Google Colab platform. Preprocessing and classification stages for machine learning takes a small amount of time, which is around 20 seconds for 1000 hyperparameter combination to have better accuracy. Deep learning model also takes 20 minutes with 93% accuracy and 64 × 64 image size without any transfer learning or pretrained model. In short, Google Colab platform is very efficient for developing machine learning models which can be developed that have high accuracy for a short time as other platform that requires any expensive hardware.

II. METHODOLOGY

In this section, algorithms and important hyperparameters selection criteria are explained in detail. 7 different machine learning algorithms such as logistic regression, kNN, SVM, Naive Bayes, Random Forest, XGBoost and a deep neural network (CNN) are presented. Hyperparameters and their effects are different for each algorithm and dataset in the model. The most used open access dataset that is Herlev dataset is selected to compare between ours and others’ methodology and hyperparameters. The dataset has 20 morphological features with its labels that are important for machine learning algorithms. Thereby, the performance of models is compared for same condition by using same features. Results of all developed models are measured by using scikit-learn library. The obtained hyperparameters and results of the models in which conditions are mentioned. Especially, computational time of training of the CNN model is reduced to 22 seconds for each epochs. The following subsections include detailed information about models and their parameters.

A. Machine Learning

Traditional machine learning algorithms are used in many areas like regression, classification. The machine learning algorithms use features of data to develop a model. At the same time, feature selection is a crucial part for machine learning, as the features must be distinctive, and changes between the measurements for each specific cell classes must be non-negligible and measurable. The distinctive features increase the performance of the model for classification with the highest metrics. Thereby, these features have significant properties of cervical cells, which are nucleus and cytoplasm area with their ratio, brightness, shortest and longest diameter, elongation, perimeter, position, maxima, minima and roundness. Logistic regression, k-Nearest Neighbours k-NN, SVM with different kernels, Naive Bayes, Decision Trees, Random Forest and XGBoost classifiers are used for classifications of normal and abnormal cervical cells by using features of the dataset.

1) Logistic Regression: Logistic regression uses the logistic function to classify data especially in statistical analysis since it helps to understand the relation between the probability of a variable and a feature of a dataset [9]. Categorical classification for this is convenient. In this study, logistic regression is used to observe the performance of the classification of papsmear images as abnormal and normal cells.

2) k-Nearest Neighbors Classifier: kNN is a supervised and lazy machine learning algorithm which is capable of memorizing the dataset. In this way, it is a robust algorithm for an old dataset [10]. However, it requires a larger memory size since, it keeps all states in calculating distances. In this study, the highest accuracy is observed by using the number of neighbors, the distance metric and the power parameter is selected 9, Minkowski, Euclidean distance respectively.

3) Support Vector Machine Classifier: SVMs are a supervised machine learning algorithm based on the statistical learning theory. SVMs are used in many areas like face, optical character, voice recognition with many advantages like high efficiency for a high dimensional dataset that the algorithm splits data with the best line [11]. Because of the advantages of kernels, the kernel of SVM is specified as radial basis function (rbf) for the highest accuracy.

4) Naive Bayes Classifier: Naive Bayes classifier is aimed to classify the data according to probability principles, the ground of Bayes Theorem. It calculates all the probabilities for each data and classifies by the highest probability. It is used in many areas like spam filtering, so Gaussian Naive Bayes classifier is considered for our work [12].

5) Decision Tree Classifier: A decision tree is a method that uses a series of decision rules to split data into smaller clusters. This approach is an old but efficient tree-based algorithm that is the fundamentals of some algorithms like random forest. For this work, the function to measure the quality of a split is based on the entropy criterion [13] also, to see entropy of the dataset.

6) Random Forest Classifier: Random forests are an advanced tree-based algorithm that choose subsets from both the features and the dataset to avoid overfitting. Its advantage is that random forest can produce a high accuracy without hyperparameter optimization. Also, the probability of overfitting is lower than decision trees. However, the number of trees in random forest is a very important parameter and it should be selected well [14].

7) XGBoost Classifier: XGBoost is a very popular and a new machine learning algorithm which is very efficient since it wins half of the challenges in machine learning platform like Kaggle [15]. XGBoost is a decision tree based and a gradient boosted algorithm. It has many systems and algorithmic optimizations like tree pruning, parallelization, and cross-validation [16].

B. Deep Learning

CNN is based on convolution integral where as in the convolution layer, which includes an activation function like Rectified Linear Unit (ReLU) to add nonlinearity to the model.
In this work, a standard CNN is used for classification and analysis of visual imagery. Image size, train-test split ratio, convolutional layer number, dropout and epoch number are some properties of CNN. The properties of neural network and the dataset that are used in this work are shown in Table I. In addition to this, the block diagram of CNN is shown in Figure 1. The model is developed with Keras with 4 convolution layer.

### Metrics

The performance of models is measured and compared with each other by using metrics. The metrics show accuracy, recall, precision, specificity and $F_1$ score that are calculated by using a confusion matrix. The confusion matrix includes the essential parameter of metrics. The metrics can be calculated by using the essential parameters according to following criterias:

- **True Positive (TP)** value is an outcome where the model predicted the dataset positive in a correct way.
- **True Negative (TN)** value describes how many negative values are predicted correctly.
- **False Positive (FP)** value represents that predicted value is positive but it is not.
- **False Negative (FN)** value represents that negative values are predicted wrongly.

These are essential metrics and other metrics like accuracy that give more information about models that are obtained by using TP, TN, FP and FN values.

#### III. Results

In this work, the performance of models is obtained for cervical cancer classification. A table is used to compare easily between different models. In addition to this, the models that have the best results are highlighted in this section. All results of the metrics of models are proposed and compared in Table II.

| TABLE I | PROPERTIES OF THE CNN |
|---------|-----------------------|
| Property | Value                |
| Image Size | (64,64,3) |
| Test Data/All Data | 0.15 |
| Convolutional Layer Number | 4 |
| Dropout | 0.4 |
| 2D Pooling Size | 2x2 |
| Validation Data/Training Data | 0.15 |
| Epoch Number | 50 |

| TABLE II | METRICS |
|----------|---------|
| Metric | Formula |
| Accuracy | $\frac{TP + TN}{TP + FP + TN + FN}$ |
| Recall | $\frac{TP + FN}{TP}$ |
| Precision | $\frac{TP}{TP + FP}$ |
| Specificity | $\frac{TN + FP}{TN}$ |
| $F_1$ Score | $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ |
TABLE III
RESULTS OF METRICS OF MACHINE AND DEEP LEARNING APPROACHES

| Metrics     | Name     | Logistic Regression | k-NN | SVM | Naive Bayes Classifier | Decision Tree | Random Forest | XGBoost Classifier | CNN for Training Data | CNN for Test Data |
|-------------|----------|---------------------|------|-----|------------------------|---------------|---------------|---------------------|-----------------------|------------------|
| Accuracy (%)|          | 83                  | 85   | 83  | 80                     | 80            | 83            | 85                  | 99                    | 93               |
| Recall (%)  |          | 83                  | 87   | 85  | 92                     | 77            | 80            | 87                  | 99                    | 93               |
| Precision (%)|         | 85                  | 86   | 83  | 76                     | 86            | 89            | 87                  | 99                    | 96               |
| Specificity (%)|        | 80                  | 83   | 83  | 87                     | 74            | 78            | 84                  | 98                    | 89               |
| F1 Score (%)|          | 85                  | 86   | 84  | 84                     | 80            | 84            | 87                  | 99                    | 95               |

III. Firstly, in machine learning techniques; logistic regression has 83% accuracy and recall value, 85% precision and F1 score with 80% specificity results. For this reason, logistic regression is an average model for Herlev dataset to classify pap smear images. kNN model has 85% accuracy, 87% recall, 86% precision and F1 score with 83% specificity and it is a good model. SVM has 83% accuracy, precision and specificity with 85% recall and 84% F1 score that is an average model. Naive Bayes classifier has 80% accuracy and 76% precision with the weakest results, also it has 92% recall and 87% specificity results with the best results. However, it is not a convenient model because of its inconsistency. Decision tree is the weakest model in machine learning algorithm with the lowest 4 metric values that are accuracy, recall, specificity and F1 score. It has 80% accuracy and F1 score, 77% recall, 86% precision and 74% specificity values. Random forest is an average model with 83% accuracy, 80% recall, 89% precision with the highest value, 78% sensitivity and 84% F1 score. XGBoost is also the best model. It has 85% accuracy and 87% F1 score with the highest values, 87% recall and precision and 84% specificity values. The best result is achieved by using XGBoost classifier with 85% accuracy. Likewise, k-NN is the second best classifier for the pap smear images specific to Herlev dataset with 85% accuracy. As concerns deep learning, despite it is standard CNN, its model has 99% accuracy, recall, precision and F1 score with 98% specificity value for training data. However, a more important thing is the model performance for test data with the model. Similarly, the model has a high accuracy for test data about 93% accuracy and recall, very good F1 score with 95% and precision with 96% and the last metric is specificity that has 89% value.

IV. CONCLUSION

In this study, machine and deep learning methods are investigated and compared to find out which model would be an appropriate choice for cervical cell classification over Herlev dataset. A model or algorithm can be suitable for a dataset. However, if the hyperparameter of a model is not optimized for the dataset, the real performance of the model will not be seen. Thus, hyperparameter optimizations and selected features play a critical role to get a good model. The results of these models indicate that the models can obtain the highest performance with hyperparameter optimization. In addition to this, the CNN model extracts its features, and it can select more distinctive features than the hand-crafted features. The CNN model has better results than traditional machine learning techniques, as it is seen in this work.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: a cancer journal for clinicians, vol. 68, no. 6, pp. 394–424, 2018.
[2] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermse, Q. F. Manson, M. Balkenhol, et al., “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” Jama, vol. 318, no. 22, pp. 2199–2210, 2017.
[3] S. S. Han, G. H. Park, W. Lim, M. S. Kim, J. I. Na, I. Park, and S. E. Chang, “Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network,” Plos one, vol. 13, no. 1, p. e0191493, 2018.
[4] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, “Pap-smear benchmark data for pattern classification,” Nature inspired Smart Information Systems (NiSIS 2005), pp. 1–9, 2005.
[5] L. Zhang, L. Lu, I. Nogues, R. M. Summers, S. Liu, and J. Yao, “Deeppap: deep convolutional networks for cervical cell classification,” IEEE journal of biomedical and health informatics, vol. 21, no. 6, pp. 1633–1643, 2017.
[6] Y. Marinakis, G. Dounias, and J. Jantzen, “Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification;,” Computers in Biology and Medicine, vol. 39, no. 1, pp. 69–78, 2009.
[7] T. Chankong, N. Theera-Umporn, and S. Auephanwiriyakul, “Automatic cervical cell segmentation and classification in pap smears;,” Computer methods and programs in biomedicine, vol. 113, no. 2, pp. 539–556, 2014.
[8] B. Ashok and P. Aruna, “Comparison of feature selection methods for diagnosis of cervical cancer using svm classifier,” Int. J. Eng. Res. Appl, vol. 6, pp. 94–99, 2016.
[9] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, Applied logistic regression, vol. 398. John Wiley & Sons, 2013.
[10] L. E. Peterson, “K-nearest neighbor;,” Scholarpedia, vol. 4, no. 2, p. 1883, 2009.
[11] C. Cortes and V. Vapnik, “Support vector machine;,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995.
[12] I. Rish et al., “An empirical study of the naive bayes classifier;,” in IJCAI 2001 workshop on empirical methods in artificial intelligence, pp. 41–46, 2001.
[13] J. R. Quinlan, “Induction of decision trees;,” Machine learning, vol. 1, no. 1, pp. 81–106, 1986.
[14] L. Breiman, “Random forests;,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.
[15] D. Nielsen, “Tree boosting with xgboost-why does xgboost win” every machine learning competition”;,” Master’s thesis, NTNU, 2016.
[16] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, “XGboost: extreme gradient boosting,” R package version 0.4-2, pp. 1–4, 2015.
[17] Y. LeCun, Y. Bengio, et al., “Convolutional networks for images, speech, and time series;,” The handbook of brain theory and neural networks, vol. 3361, no. 10, p. 1995, 1995.