ECG Heartbeat Classification

Yaozhuo Wang
ECE
Department
Western
University
London, Canada
ywan3223@uwo.ca

***Abstract*—Cardiovascular disease is known as one of the leading causes of death worldwide. The Electrocardiogram (ECG) data, the measured electrical signal of the heart, is a reliable and effective measure to detect problems in the cardiovascular system. In this paper we propose a classification model approach based on ResNet and transfer learning techniques. The method uses the MIT-BIH Arrhythmia Dataset and the PTB Diagnostic ECG Database, which classifies five different arrhythmias by diagnostic type. The imbalanced data was eliminated by using SMOTE and ENN. The model was optimized using a Cyclical Learning Rate and Grid Search. The model was optimized to achieve 99.743% accuracy in the prediction of arrhythmias.**

***Keywords*—ECG, ResNet, transfer learning, cardiovascular disease prediction**

I. Introduction

Cardiovascular disease has become the number one killer of patients worldwide, with more people dying from cardiovascular disease than any other cause of death each year. In 2016, approximately 17.9 million people died from cardiovascular disease accounting for 31% of all deaths worldwide. Of these, 82% occurred in middle-income countries [1] and cardiovascular health is a growing concern. ECG data is a presented as a continuous wave wtih each segment relating to different time periods of a heart's beating cycle. In clinical trials, the central electrical signal does not appear smoothly or as predicted. Manual recognition of the signal has the disadvantage of a high rate of misdiagnosis and untimely diagnosis, while cardiovascular disease is characterized by rapid onset and needs early treatment. Therefore, accurate classification of ECGs has become an important issue in the medical field. As a result, more artificial intelligence is now being used to classify ECGs, making them more accurate and less costly.

In this paper our project uses machine learning to classify ECGs more accurately to detect and inform patients of heart rate abnormalities. At the beginning of the project, we used the SMOTE-ENN technique to eliminate imbalance in our data to avoid deceptively high accuracy in the classification results.

The Residual Network (ResNet) was used to structure the ECG analysis network and map the collected feature data into a one-dimensional vector. We then tune the ResNet model to optimize the results. Lastly, we implement transfer learning to reuse the pre-trained model weights for another arrythmia dataset.

The remaining components of the paper are listed below.

II. Dataset

III. Background

IV. Related Work

V. Methodology

VI. Results

VII. Conclusion

II. Dataset

Two collections of heartbeat signals were derived from two famous heartbeat classification datasets, the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) Arrhythmia Dataset and the Physikalisch-Technische Bundesanstalt (PTB) Diagnostic ECG Database. The MIT-BIH database is a large publicly available database containing samples retrieved by the BIH Arrythmia Laboratory between 1975 and 1979; these samples are mainly used for investigations in heart arrythmia detection [2]. The PTB Diagnostic ECG dataset is a collection of ECG data collected from healthy volunteers and patients with different diseases for the usage in research, algorithmic benchmarking, or teaching purposes [3].

A. Massachusetts Institude of Technology-Beth Israel Hospital (MIT-BIH) Arrhythmia Dataset

Originating from Boston's Beth Israel Hospital, this dataset contains 48 half-hour excerpts of two-channel ECG data recordings from 47 subjects. 23 of which come from a random sample of a larger 4000 dataset of 24-hour ambulatory ECG recordings; this random sample contains a mixed population of 40:60 inpatient to outpatient within the hospital. The remaining 25 samples were hand-picked in order to express clinically significant arrythmias within the dataset [2].

In total, the dataset contains 109446 samples with 188 discrete time values within each sample representing the heart rate of a patient. The total samples provided by the database were split into 87943 for training and 21503 for test data. There contains a total of 5 possible category labels associated with each sample; N, normal beat; S, supraventricular premature; V, premature ventricular contraction; F, fusion of ventricular and normal beat; Q, unclassified beat. Table I outlines the detailed annotations of each category.

TABLE I. MIT-BIH CATEGORY LABELS AND ANNOTATIONS [4].

Category	*Annotations*
N	Normal, Left/Right bundle branch block, Atrial escape, Nodal escape
S	Atrial premature, Aberrant atrial premature, Nodal premature, Supra-ventricular premature
V	Premature ventricular contraction, Ventricular escape
F	Fusion of ventricular and normal
Q	Paced, Fusion of paced and normal, Unclassifiable

B. Physikalisch-Technische Bundesanstalt (PTB) Diagnostic ECG Database

The ECG collection of this dataset came from the Department of Cardiology of University Clinic Benjamin Franklin in Berlin, Germany; PTB, or the National Metrology Institute of Germany, then compiled and distributed the digitized ECG collection. The dataset is comprised of 290 subjects (aged 17 to 87, mean 57.2; 209 men, mean age 55.5, and 81 women, mean age 61.6; ages were not recorded for 1 female and 14 male subjects) [3].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

In total, the dataset contains 14552 samples with a similar 188 discrete time values for each sample. There contain only two classification categories within the dataset; normal and abnormal. This is split into 4046 and 10506 samples respectively.

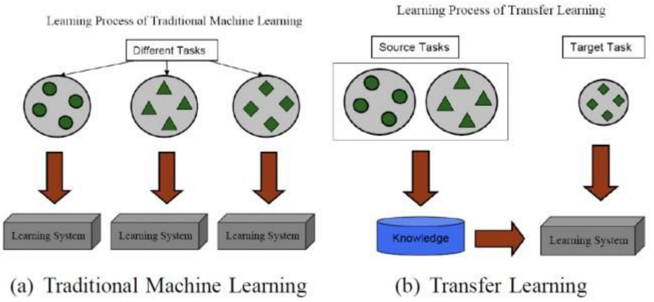
III. Background

A. Network Blocks

a) Residual Networks (ResNet)

Residual Networks, or ResNets are made up of residual blocks. Residual blocks have skip connections between the layers of a model which add the outputs from the previous layers to the output of the stacked layer. This residual block adopts a shortcut connection which allows it to directly learn residuals for the target output. Thus, we avoid the problem of vanishing/explode gradients due to the huge number of convolutional and max pooling layers. ResNets help optimize the neural network easier by ultimately making a simpler network [5].

b) Transfer Learning



(a) Traditional Machine Learning (b) Transfer Learning

Fig. 1. Different learning processes between (a) traditional machine learning and (b) transfer learning [6].

The assumption for most machine learning problems is that the training data and test data lie in the same feature space and have the same mathematical distribution. However, in most practical applications, this assumption may not hold. Figure 1 illustrates the difference between traditional machine learning and transfer learning.

Especially in the field of biomedicine, researchers need to solve multi-label classification problems. However, there are rarely a large number of labeled samples in the current medical field. Transfer learning is the idea of overcoming the isolated learning paradigm and utilizing knowledge acquired for one task to solve related ones.

B. Hyperparameters

Hyperparameters are variables that define a model's architecture and affect the final classification result of the neural network. Hyperparameter tuning is the process which the machine automatically explores and selects the best model of the architecture [7]. Here, three hyperparameters will be discussed; learning rate, batch size, epoch, and momentum.

Learning rate is the most crucial hyperparameter for training. Cyclical learning rate determines the convergence or divergence of the learning rate in an easy way [8]. Batch size is a factor for descending gradient that drives the number of training samples to be processed before updating the internal parameters of the model [8]. Epoch determines how many times the learning algorithm will run over the whole training dataset. In each epoch, all the data will be used exactly once and made up of one or more batches [8]. The momentum determines how much the most recent update can affect the most recent weight update. It is used to make the algorithm keep the direction along for some time before it changes direction [9].

C. Evaluation Metrics

Four metrics were used to evaluate the multi-class and binary classification dataset. These metrics are appropriate for classification as they are based on correct or incorrect predictions compared to the test data.

A confusion matrix is a graphic representation of the true and false classifications between the test set and the predicted set of outcomes. Predictions are categorized as True-Positive, True-Negative, False-Positive, and False-Negative.

Accuracy considers each of these classes from the confusion matrix and determines the percentage of correct classifications.

Precision calculates the ratio of true positive classifications to true positive and false positive classifications. Recall calculates the ratio of true positive calculations to true positive and false negative classifications.

IV. Related work

ECG classification is a task with high demand due to its importance in diagnostics within the medical field. Medical and computer science researchers alike have tackled this problem with most recent solutions revolving around machine learning (ML) and deep learning (DL). While ML algorithms can be quite robust, they are reported to only capture shallow feature learning [10]. DL architectures can overcome this issue with large neural networks for more complex and deeper learning. A successful convolutional neural network (CNN) containing 34-layers have been able to outperform cardiologists in both recall and precision (80% vs. 72.3% and 78.4% vs. 72.4% respectively) [11]. Another successful, and different, DL architecture involved a recurrent neural network (RNN) to analyze the time correlation of ECG signals. Experimental results provided an accuracy of 98.7% and 99.4% for supraventricular premature (S) and ventricular contraction (V) classification respectively [12]. The inspiration for this paper will follow a residual CNN architecture utilizing the same datasets as was discussed to allow for direct comparison. Evaluation of their proposed model suggested an accuracy of 93.4% [13] and will used as the main comparator.

V. Methodology

A. Data Preprocessing

The important first steps of cleaning and parsing training and testing data is integral for proper training and evaluation of the model. The exact aspects to be explored include the overall large quantity of data and the imbalance of categories within it. Imbalanced data can cause overfitting to category labels with the largest quantity; this is referred to as the majority class. Large quantities of data may be good for robust training of the model but has the trade-off of increased run-times that may not be necessary, i.e., diminishing returns.

a) Imbalanced Data

Focusing on a binary classification example, an imbalanced dataset is one where one of its two class labels has a significantly higher quantity count than the other; these will be called the majority and minority class respectively. As previously mentioned, a model trained with imbalanced data will learn many features about the majority class but may not entirely be represented by the minority class. This could also be not as obvious to see if the test dataset is similarly imbalanced, where the high accuracy would not be indicative of the minority class features.

The most common algorithm to handle rebalancing of the data is the Synthetic Minority Oversampling Technique (SMOTE). This algorithm uses a K-nearest neighbours approach where each minority feature will find the K closest feature to it and generate a new synthetic data-point at a random distance between them. The main aim of this technique is to generate a 1:1 proportion of majority to minority data points.

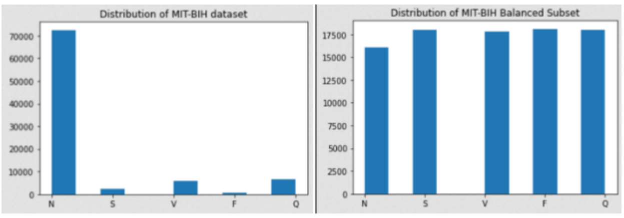
To increase generalization and robustness, undersampling is also important to consider. Edited Nearest Neighbours (ENN) is one such undersampling algorithm that aims to clean-up any data points that may cause misclassifications within the hyperplane. Using a similar K-nearest neighbours approach, ENN will find features within the wrong margins and boundaries of the hyperplane. This will allow for a cleaner separation of the data and increase the classification's performance.

With both algorithms used, a hybrid algorithm SMOTE-ENN will allow for balancing of data to prevent over-fitting in trade of a larger quantity dataset and run-time. To extend this algorithm from binary to multi-class classification, the total multi-class data balancing will be broken down into multiple binary data balancing where each combination of category pairs will undergo SMOTE-ENN.

b) Data Subset

With the unaltered MIT-BIH training dataset containing 87943 samples (where the majority class is the N category with a total of 72471 samples), post-balancing will bring the total dataset to over 350000 samples. This is quite a large dataset and would require a very powerful computational unit and time to process. To save run-time, it would be wise to take a subset of this balanced dataset. As the SMOTE-ENN itself takes some time, it would be advantageous to first take a subset of the minority class and then perform balancing. This will not only save run-time in training the model, but also save run-time for balancing as now the 1:1 proportion that SMOTE-ENN will require to match is of a lower quantity.

A 25% subset was taken for the majority class (category N) and then balanced to produce the distribution as shown in Fig. 2.



(a) (b)

Fig. 2. MIT-BIH dataset distribution before (a) and after (b) preprocessing

Finally, for both the binary and multi-classification datasets, the hold-out method was applied. 64% of the data was used for training, 16% for validation, and 20% was used for testing.

B. Network Architecture

a) ResNet

The model is comprised of ResNet residual block with the ECG data points as input and target class categories as outputs By using the ResNet residual block, as shown in Figure 3, ResNet-18 can be used as model for classifying ECG signals.

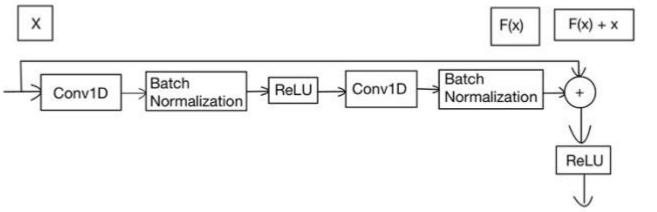


Fig. 3. The ResNet residual block [14]

For the input ECG signal data, a one-dimension convolution kernel of length 32 is used to extract features from the input data. There will be batch Normalization, ReLU and max pooling followed by the convolution. In image classification, ResNet18 always uses a 3*3 convolution kernel, but since the ECG signal is different from image data, the few sample points at a given time-period cannot produce a change. Therefore, a large convolution kernel is used here to solve this problem.

Figure 4 shows the architecture of the ResNet-18 model, including three parts: convolutional layer, ResNet-18 residual block layer, and fully connected layer. In the first phase of our model, the convolution layer is used to do low level feature extraction on the input data in preparation for the next level of processing. The second phase uses ResNet-18 residual block. In this section, the data in each residual block will go through the convolution layer twice and the batch normalization process, ReLU, is added between the two convolutions. Finally, the original data and the processed data of the same size are added together to complete the block module creation. The purpose of this step is to inherit the optimal efficiency from the previous step and make the model continue to converge.

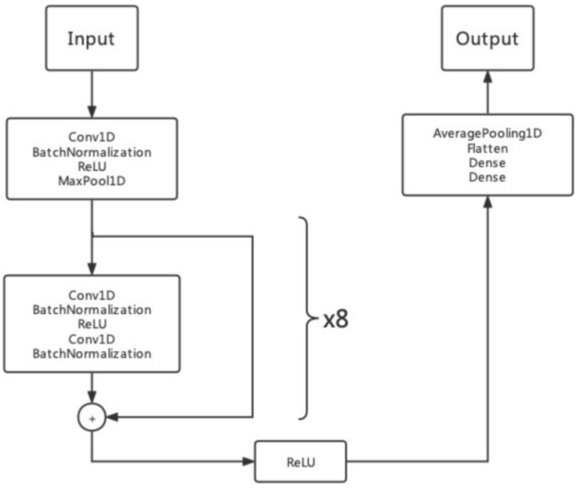


Fig. 4. The architecture of ResNet18

For better performance, we add two batch norms in each residual block to speed up the neural network training and convergence. The model goes through this residual block eight times before the data is sent to the last phase, which is the fully connected layer.

The features collected in the first two stages are mapped to a one-dimensional vector and the vector is regressed using the SoftMax [15]/sigmoid function, which is suitable for the multi-objective classifier/binary classifier. These two classifications represent different models for the two data sets.

b) Transfer Learning

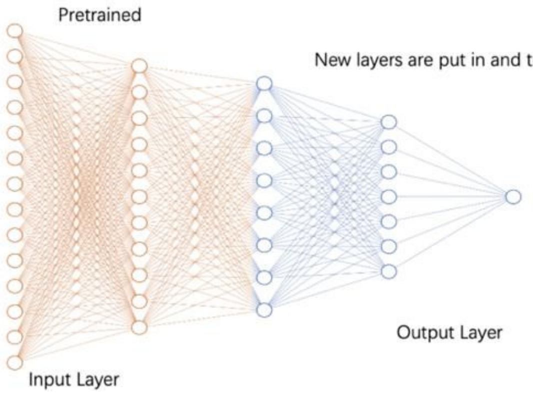


Fig. 5. Transfer learning model

In the previous step, we have trained deep Residual neural network (Resnet) on both the MIT-BIH and PTBD data set. The weights of the two models have been stored locally.

After that, the weight of the MIT-BIH model can be set to the untrained PTBD model in addition to the last three full connection layers of the model will be removed. To increase the training speed, all weights of the pre-trained model will be frozen. Since the MIT-BIH dataset has five categories, but the PTB dataset is only has a binary categories, we need to re-build a new full connection layer to adapt to the mismatching outputs After building the model and setting the weights, the training set of PTB dataset can be provided for model training. The transfer learning model can be found in Figure 5.

C. Model Tuning

Two methods were used to tune the hyperparameters, these are Cyclical Learning Rate and Grid Search. Tuning hyperparameter can only make small changes on the accuracy of the model, which is important in a medical-related topic where every additional point of accuracy counts.

a) Cyclical Learning Rate

To use the Cyclical learning rate, one must provide a step size, as well as minimum and maximum learning rate boundaries. The step size is the quantity of epochs used for each step. A cycle has two steps which are half cycles of the iteration. One is a linear increase in learning rate from the minimum to the maximum, whereas the other is a linear decrease.[8]

LRFinder is used to find the range of the learning rate. Starting with a very low learning rate in an epoch and changing it in each small batch size by multiplying by some factor until it reaches a very high value. Record the losses The losses are recorded for each iteration. And when Once its finished, these losses will be compared with the learning rate. Figure 7 shows the relation of loss and learning rate.

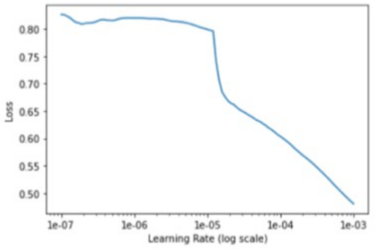


Fig. 6. Range of learning rate.

b) Grid Search

Model optimization is carried out using grid-based hyperparameter tuning. The three hyperparameters to undergo such are the momentum, epoch, and batch size. Table II shows the exact parameter values used to tune the ResNet model. Each set of hyperparameters is used to train the model via the grid search algorithm, which chooses the set of hyperparameters that has the lowest validation set error.[16] The hyperparameters considered are in table 2.

TABLE II. TUNING HYPERPARAMETER

hyperparameters	***value***
Batch Size	64, 128, 256, 512
Epoch	25, 50, 75
Momentum	0.2, 0.4, 0.6, 0.8, 0.9

VI. RESULTS

For this application of arrythmia classification, it is most important to minimize false negatives, where the patient is diagnosed ‘normal’ even though they have an ‘abnormal’ result. Doing so will maximize recall. If a patient were to receive a false negative diagnosis, they would likely not seek further testing to later disprove the diagnosis, which is ultimately a health concern for a patient.

A. Residual Network - Multiclass Dataset

The results before tuning the ResNet models are shown in Figure 7 below. An accuracy of 99.681%, with a precision of 99.662% and recall of 98.903% was achieved, respectively. The grid search is applied to pursue minimizing the false negatives, noted in red in figure 7c).

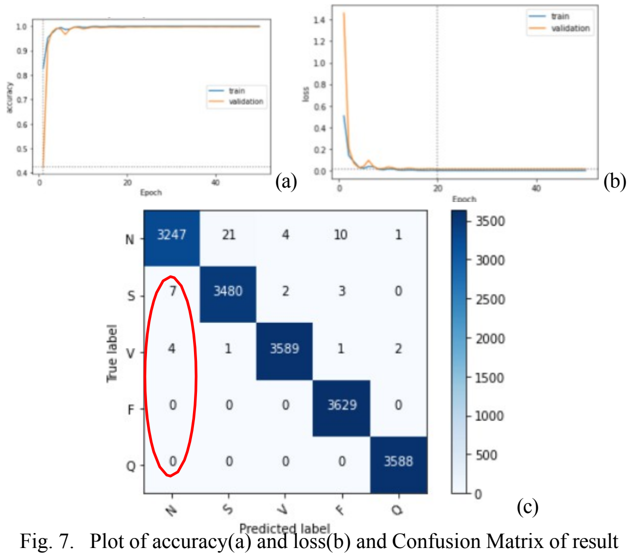


Fig. 7. Plot of accuracy(a) and loss(b) and Confusion Matrix of result before tuning(c)

The resulting best parameters as identified by grid search is as follows: Batch Size = 64, Epochs = 75, Momentum = 0.9. The results after tuning the ResNet models are shown in Figure 8 below. An accuracy of 99.676%, with a precision of 99.571%, and recall of 99.056% was achieved, respectively. The number of false negatives did increase (11 → 14), and consequently the recall decreased slightly. The grid search did recommend a higher epoch than the original model, which may have caused overfitting. These results show that the grid search model is less effective than the pre-grid search model. Thus, the original model is used for transfer learning.

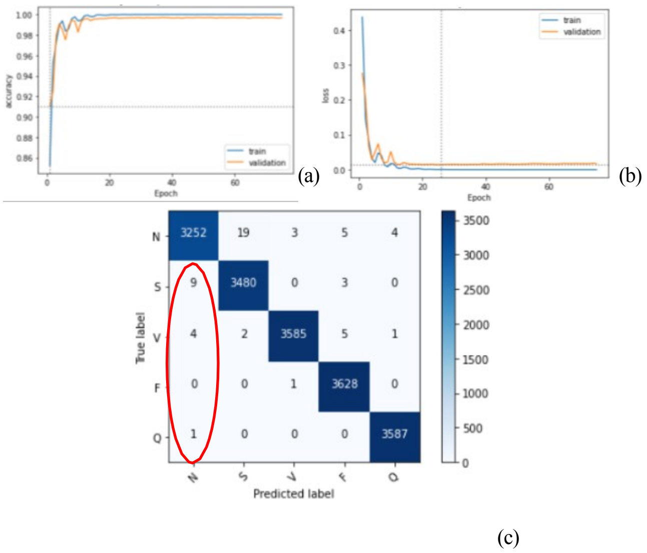


Fig. 8. Plot of accuracy(a) and loss(b) and Confusion Matrix of result after tuning(c)

B. *Residual Network - Binary Dataset*

The results before tuning the ResNet models are shown in Figure 9 below. An accuracy of 99.692%, with both a precision and recall of 99.703%.

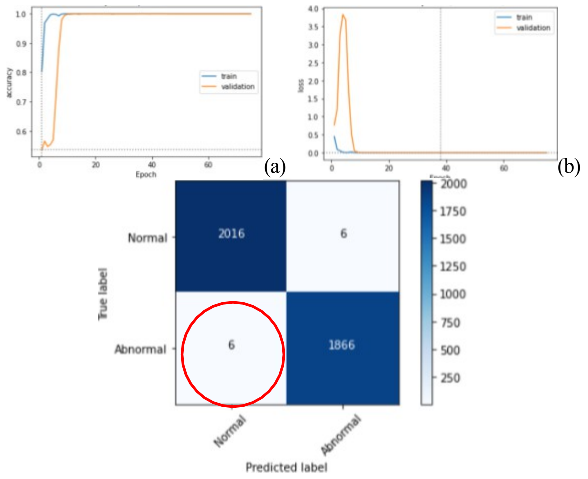


Fig. 9. Plot of accuracy(a) and loss(b) and Confusion Matrix of result before tuning(c)

The resulting best parameters as identified by grid search is as follows: Batch Size = 64, Epochs = 50, Momentum = 0.9. The results after tuning the ResNet models are shown in Figure 10 below. An accuracy of 99.743%, with a precision of 99.627%, and recall of 99.840% was achieved, respectively. This increase in recall corresponds to a decrease in false negatives (6→3). This grid search successfully maximizes recall of the model and is implemented in transfer learning.

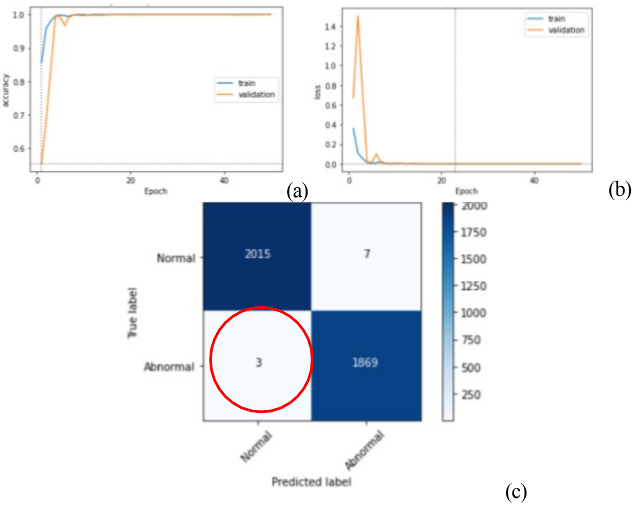


Fig. 10. Plot of accuracy(a) and loss(b) and Confusion Matrix of result after tuning(c)

C. *Transfer Learning*

The results of transferring the multi-class classification weights to the binary classification dataset are shown in Figure 11 below. An accuracy of 99.538%, precision of 99.459%, and recall of 99.567% are achieved. While the accuracy is overall lower, the transfer learning does still achieve minimizing the false negatives. This result indicates the model is not overfitted.

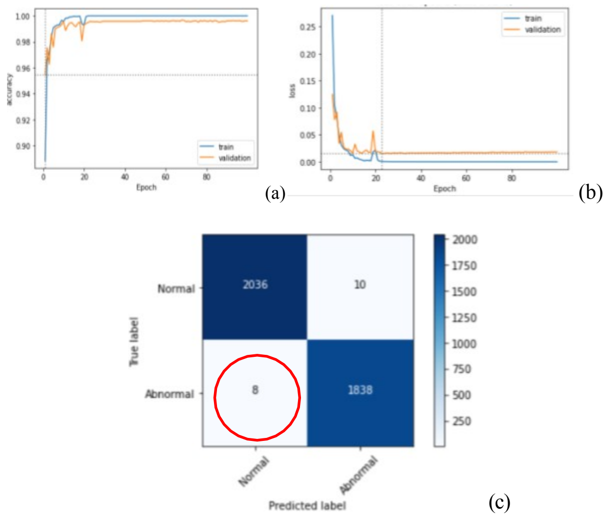


Fig. 11. Plot of accuracy(a) and loss(b) and Confusion Matrix of result after transfer learning(c)

The results of transferring the binary weights to the multi-class classification dataset are shown in Figure 12 below. An accuracy of 98.618%, precision of 98.610%, and recall of 97.259% are achieved. The accuracy and loss curves do follow the anticipated trendline but have an oscillatory nature. This behavior could indicate the model is overfitted, as it has incorrectly learned random fluctuations in the training data. This prediction is verified by the poor recall in comparison to all other models.

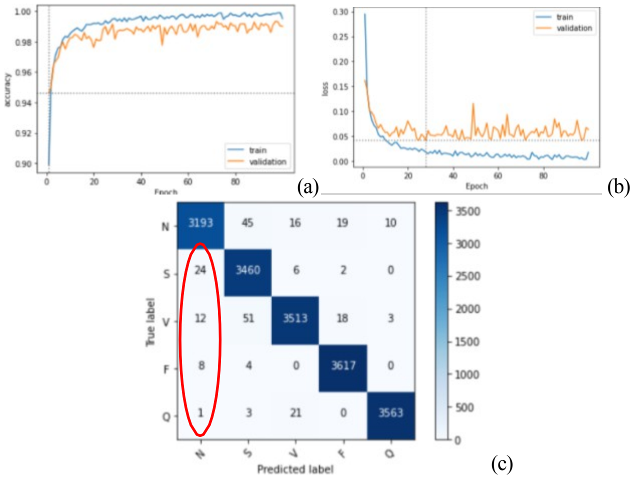


Fig. 12. Plot of accuracy(a) and loss(b) and Confusion Matrix of result after transfer learning to multiclass classification(c)

VII. Conclusion

This paper presented ECG classification results with the method using a ResNet model architecture combined with transfer learning. Specifically, a deep convolutional neural network has been trained with a residual block for arrhythmia classification. The model classifies five different arrhythmias as well as a numerical control classification by diagnostic category. The best resulted accuracy on validation data is 99.743%. This accuracy is significantly higher than the related ResNet model with a 93.4% accuracy [13].

In the future, it is possible to enhance the grid search by increasing the number of hyperparameters. Doing so would identify other hyperparameters that can increase recall. The overfitting problem can be solved by implementing the drop-out method to stop model training earlier. In any case, the fundamental objective is to present and apply a deep learning method that competes in classification accuracy with human findings without relying on feature engineering.

References

1. “Cardiovascular diseases (cvds),” *World Health Organization*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed: 07-Dec-2022].
2. G. Moody and R. Mark, “MIT-BiH Arrhythmia Database,” *MIT-BIH Arrhythmia Database v1.0.0*, 24-Feb-2005. [Online]. Available: https://www.physionet.org/content/mitdb/1.0.0/. [Accessed: 05-Dec-2022].
3. R.-D. Bousseljot, “PTB diagnostic ECG Database,” *PTB Diagnostic ECG Database v1.0.0*, 25-Sep-2004. [Online]. Available: https://www.physionet.org/content/ptbdb/1.0.0/. [Accessed: 05-Dec- 2022].
4. M. Kachuee, S. Fazeli, and M. Sarrafzadeh, “ECG Heartbeat Classification: A Deep Transferable Representation,” Jul. 2018. [Online]. Available: https://arxiv.org/pdf/1805.00794.pdf. [Accessed: 01-Dec-2022].
5. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
6. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.
7. L. N. Smith, “A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay,” *arXiv.org*, 24-Apr-2018. [Online]. Available: https://arxiv.org/abs/1803.09820. [Accessed: 07-Dec-2022].
8. L. N. Smith, "Cyclical Learning Rates for Training Neural Networks," 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 464-472, doi: 10.1109/WACV.2017.58.
9. M. Claesen and B. De Moor, “Hyperparameter search in machine learning,” *arXiv.org*, 06-Apr-2015. [Online]. Available: https://arxiv.org/abs/1502.02127. [Accessed: 07-Dec-2022].
10. “Deep learning for ECG classification - IOPscience - Institute of Physics.”[Online].Available: https://iopscience.iop.org/article/10.1088/1742-6596/913/1/012004. [Accessed: 17-Nov-2022].
11. Hannun AY;Rajpurkar P;Haghpanahi M;Tison GH;Bourn C;Turakhia MP;Ng AY; “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nature medicine*. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30617320/. [Accessed: 17-Nov- 2022].
12. C. Zhang, G. Wang, J. Zhao, P. Gao, J. Lin and H. Yang, "Patient-specific ECG classification based on recurrent neural networks and clustering technique," *2017 13th IASTED International Conference on Biomedical Engineering (BioMed)*, 2017, pp. 63-67, doi: 10.2316/P.2017.852-029. [Accessed: 17-Nov-2022]
13. M. Kachuee, S. Fazeli, and M. Sarrafzadeh, “ECG Heartbeat Classification: A deep transferable representation,” *arXiv.org*, 12-Jul-2018. [Online]. Available: https://arxiv.org/abs/1805.00794. [Accessed: 17-Nov-2022].
14. Y. Lecun and Y. Bengio, “Convolutional networks for images, speech, and time-series, ” Handbook of Brain Theory & Neural Networks, 1995.
15. F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin Softmax for face verification,” IEEE Signal Processing Letters, vol. 25, no. 7, pp. 926 –930, 2018.
16. B. H. Shekar and G. Dagnew, "Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data," 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), 2019, pp. 1-8, doi: 10.1109/ICACCP.2019.8882943.