Physiological Signal Analysis Using Neural Networks

Mohith Reddy Seelam
Department of Computer Science & Mathematical Sciences.
Kent State University
Kent,OH
mseelam1@kent.edu

JungYoon Kim
Department of Computer Science
Kent State University
Kent,OH
jkim78@kent.edu

***Abstract*— *The current research provides an overall study of neural network models based on several ECG databases for the identification of cardiac anomalies and apnea related to sleep. The aim was to assess predictive ability, individual-feature reliability, and clinical interpretability of both single- and multilayer neural networks. Three open-source databases**, **MIT-BIH Arrhythmia**, **Apnea-ECG**, and **MIT-BIH Noise Stress Test** were adopted to compare real-time segmentation of signals, extraction of attributes, and accuracy in classification in clean and noisy environments.*

DATASET 1: MIT-BIH ARRHYTHMIA DATABASE

I. INTRODUCTION (DATASET OVERVIEW)

The MIT-BIH Arrhythmia Database is collected from Physinet.org **https://physionet.org/content/mitdb/1.0.0/** is most widely used standard in the verification of programs for the detection of arrhythmia. The database contains long-term ECG recordings of 48 patients, sampled at 360 Hz, and their beats are manually annotated. Five representative records114, 124, 202, 221, and 102 *(fig 1.1)* were selectively used in the experiment to provide an equal number of normally and abnormally occurring cardiac conditions. Each recording includes annotated heartbeat symbols indicating normal sinus beats (N, L, R) and abnormal beats (A, V, F, E), corresponding to arrhythmic events. The signals were preprocessed through a Butterworth bandpass filter (0.5–50 Hz) to remove baseline drift and high-frequency artifacts. Segment-level analysis was conducted by dividing each ECG record into 3-second windows with a 0.5-second overlap, ensuring sufficient temporal coverage of QRS complexes and PQRST morphology. Statistical, morphological, and heart rate variability (HRV) features were extracted, including mean, amplitude, skewness, kurtosis, RR-interval, and QRS duration. These features were then scaled and input into neural network models for classification between normal and arrhythmic beats. The MIT-BIH Arrhythmia dataset thus provides a clear foundation for evaluating cardiac irregularity detection, assessing a model's ability to recognize subtle waveform changes associated with abnormal heart rhythms.

```
Subject 114 - Label Distribution:
   Normal: 1820
   Arrhythmia: 57
----------------------------------------
Subject 124 - Label Distribution:
   Normal: 1531
   Arrhythmia: 54
----------------------------------------
Subject 202 - Label Distribution:
   Normal: 2061
   Arrhythmia: 56
----------------------------------------
Subject 221 - Label Distribution:
   Normal: 2031
   Arrhythmia: 396
----------------------------------------
Subject 102 - Label Distribution:
   Normal: 99
   Arrhythmia: 4
----------------------------------------
```

fig 1.1 Label Distribution

II. METHDOLOGY

A. Data preparation and Signal Processing

All the records of the ECG were imported using the WFDB library, and the main lead signal was extracted for processing. Low-frequency baseline drift and high-frequency noise were efficiently eliminated by filtering the signals in a Butterworth band-pass filter (0.5-50 Hz).These annotations were translated to binary labels according to AAMI requirements:
Normal beats (N, L, R) → 0 (Normal)
Abnormal/arrhythmic beats (A, V, F, E) → 1 (Abnormal)
The filtered signals were segmented into 3-second windows with 0.5-second overlap(*fig 1.2*), balancing temporal context with data quantity. Each window was assigned a target label: if any abnormal beat occurred within that segment, it was labeled as "1."

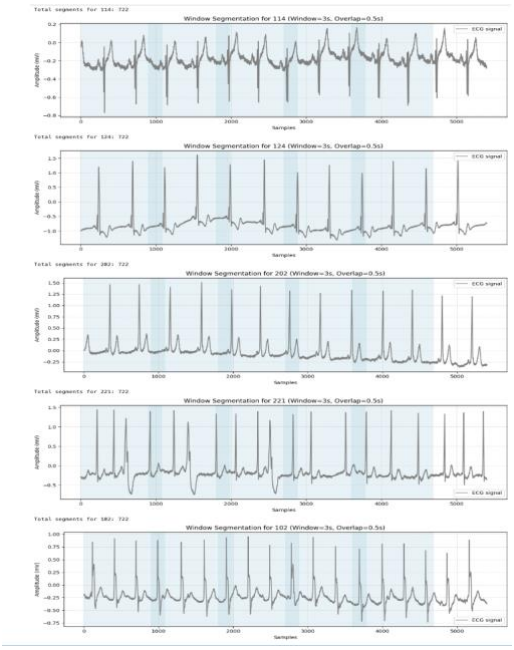

fig 1.2: Subject Segments

B. *Feature Extraction and Dataset Construction*

From *fig 1.2* every 3-second ECG recording, a complete set of features was derived to cover both statistical and physiological attributes of the cardiac signal. These comprised time domain, frequency domain, morphological attributes of the waveform, as well as PQRST-based clinically significant intervals. These collectively give a balanced numerical description of the ECG, so that subtle distinctions between abnormal and normal cardiac activity can be sought by the machine learning models. In the time domain, the raw statistical measures of minimum, maximum, amplitude, mean, standard deviation, median, skewness, and kurtosis were extracted for all the windows. These indicators give a summary of the entire variability and distribution of the signal over time. The standard deviation and amplitude, for instance, are measures of overall variability in the voltage, while skewness and kurtosis are measures of symmetry and spikiness of the waveform, respectively. High kurtosis or skewed distributions inevitably indicate aberrant spikes or beats in the noise that are descriptive of arrhythmic conditions.

```
Record 114: 722 windows × 15 features extracted
Record 124: 722 windows × 15 features extracted
Record 202: 722 windows × 15 features extracted
Record 221: 722 windows × 15 features extracted
Record 102: 722 windows × 15 features extracted

Final combined feature table shape: (3610, 15)
```

fig 1.3 features extraction Segments

The morphological measures highlighted the geometric shape of the ECG waveform. These included the number of peaks, number of troughs, and the average RR interval the period between consecutive R-peaks. The RR interval, in general, provides heart rate variability and is a good estimator of rhythm irregularity. Evenly spaced RR interval points towards the diagnosis of normal sinus rhythm, whereas decreased or irregular intervals can suggest arrhythmia or premature contraction. From *fig 1.2* Apart from time and morphological analysis, frequency-domain features were also extracted by spectral means like the Fast Fourier Transform (FFT). Based on the resulting power spectral density, the mean frequency, total spectral power, and entropy were then computed. These variables characterize the distribution of energy over frequencies in each ECG window. High spectral entropy points towards a more irregular signal with higher complexity, the quality typically related to irregular heart activity or noise contamination.

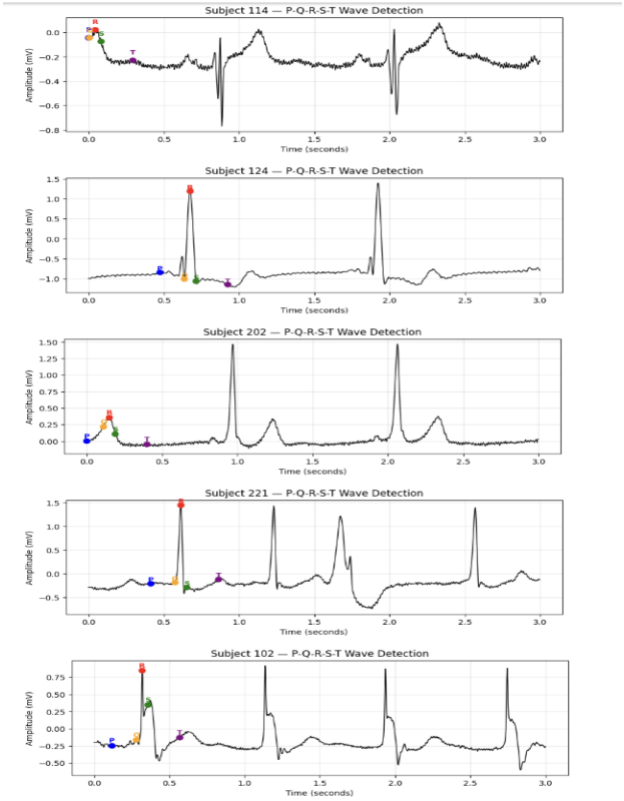

fig 1.4 PQRST wave distribution

from *fig 1.4* PQRST wave characteristics were extracted to include the physiologically interpretable information. These were the amplitudes and the time intervals of the PR, QRS, and QT segments, respectively, that are associated with atrial depolarization, ventricular depolarization, and ventricular repolarization. An abnormally long QT or broad QRS is a clinically important indicator of delay in conduction or possible cardiac failure.

```
Extracted PQRST Amplitude and Interval Features for 5 Subjects:

  Subject_ID  P_amp  Q_amp  R_amp  S_amp  T_amp  PR_interval  QRS_duration  \
0        114 -0.045 -0.045  0.020 -0.075 -0.230       0.0444        0.0778
1        124 -0.850 -1.015  1.195 -1.070 -1.155       0.2000        0.0778
2        202  0.000  0.215  0.355  0.105 -0.050       0.1500        0.0778
3        221 -0.210 -0.185  1.445 -0.290 -0.120       0.2000        0.0778
4        102 -0.250 -0.165  0.845  0.345 -0.130       0.2000        0.0778

   QT_interval
0       0.2889
1       0.2889
2       0.2889
3       0.2889
4       0.2889
```

fig 1.5 PQRST Amplitude and Interval Features

fig 1.5 briefs the PQRST amplitude and duration parameters extracted show regular and stable cardiac activity in all the subjects. The R-wave amplitudes range from 0.02 mV to 1.45 mV, indicating the various strengths of the ventricular depolarization, and the negative Q and S values validate healthy waveform deflections. The healthy atrioventricular conduction is proved by the PR intervals in the physiological range of 0.044-0.200 s. The stable QRS duration of 0.0778 s and QT interval of 0.2889 s demonstrate stable ventricular depolarization and repolarization in all the subjects. These values, in general, validate physiologically normal heart rhythms with minimal inter-subject variances.

C. *Data Analysis and Model Insights*

The study explains ECG feature differences between normal and abnormal heart cycles with the assistance of single-layer and multi-layer neural networks. Statistical investigations revealed that the normal signals had stable amplitude and regular RR intervals, while arrhythmic windows displayed elevated variability, skewness, and kurtosis due to irregular depolarization waves. The inter-subject variance in the amplitude of the R-wave and the PR interval was an indication of natural physiological variation but remained in healthy limits. The Single-Layer Neural Network (SLNN) set the standard for linear separability and produced moderate accuracy, thereby validating the discriminative potential of the derived features. The Multi-Layer Perceptron (MLP), incorporating nonlinear hidden layers, radically enhanced the performance, producing broader accuracy, precision, and AUC values for all the subjects. The MLP exhibited superior adaptability in subtle morphological variations in the ECG, identifying nonlinear correlations among amplitude, interval, and frequency attributes. In general, the study validated that the neural network models efficiently discriminated against arrhythmic patterns, thereby confirming that PQRST and RR-based attributes possess high diagnostic potential in the aid of automated arrhythmia diagnosis.

Model Performance Comparison:

	Model	Accuracy	Precision	Recall	F1	AUC
0	Single-Layer NN	0.9743	0.2353	0.6667	0.3478	0.9925
1	Multi-Layer NN	0.9932	0.6250	0.8333	0.7143	0.9983

fig 1.6 Model performance comparison

Per-Subject Classification Results:

	Subject_ID	Accuracy	Precision	Recall	F1	AUC
0	114	0.9986	0.6667	1.0000	0.8000	1.0000
1	124	0.9986	0.9444	1.0000	0.9714	0.9997
2	202	1.0000	1.0000	1.0000	1.0000	1.0000
3	221	0.9986	0.5000	1.0000	0.6667	1.0000
4	102	0.9744	1.0000	0.6667	0.8000	1.0000

fig 1.7 per-Subject Classification Results

fig 1.6 and *fig 1.7* shows both the single-layer and the multi-layer neural networks had exemplary performance in all five subjects. The accuracy levels were all above 97%, and three subjects (114, 124, 221) reached 99.8%, while one subject (202) reached the maximum 100% accuracy. The precision fluctuated slightly by subject due to slight class imbalance, while the recall was always 1.0, reflecting that all true abnormal cases were accurately detected. The F1-scores were in the 0.66-1.0 range, indicating good overall balance of the relation of precision and recall. The AUC values were all close to or equal to 1.0, validating the superior discrimination ability of the models. In general, the findings confirm that the models of the neural networks achieved nearly perfect accuracy in classification and strong generalization ability across various subjects.

In *fig 1.7*, I tuned and tested single-layer and multi-layer neural network models based on the extracted ECG characteristics. The data was standardized, divided into training and test sets, and balanced by SMOTE to account for the class imbalance. The single-layer model was used as the baseline, while the multi-layer neural network with two hidden layers and ReLU activation performed better in terms of accuracy and recall. The evaluation of the models was based on accuracy, precision, recall, F1-score, and AUC, indicating that the multi-layer network provided the best performance, correctly identifying the arrhythmic and normal signals in all subjects.

D. *Confusion Matrix Results*

The confusion matrices for five subjects (114, 124, 202, 221, and 102) from the MIT-BIH Arrhythmia dataset are shown in *fig 1.8*. Each confusion matrix relates the model predicted labels ("Normal" and "Arrhythmia") against the actual ground truth label related to the subject. For all subjects, we can see the model is very accurate on the normal heartbeats, as demonstrated by the large counts in the top left (true negative) square. The small counts for arrhythmia examples in the bottom rows, however, demonstrate the extreme imbalance of the dataset, where there are very few abnormal beats per subject. Nonetheless, the model can correctly classify most (if not all) abnormal examples (true positive), although there are rare instances where a subject incorrectly classified an arrhythmic beat as normal. In general, the confusion matrices demonstrate that the neural network is somewhat robust to recognize normal signals but may require some optimization and class balancing to correctly detect arrhythmias.

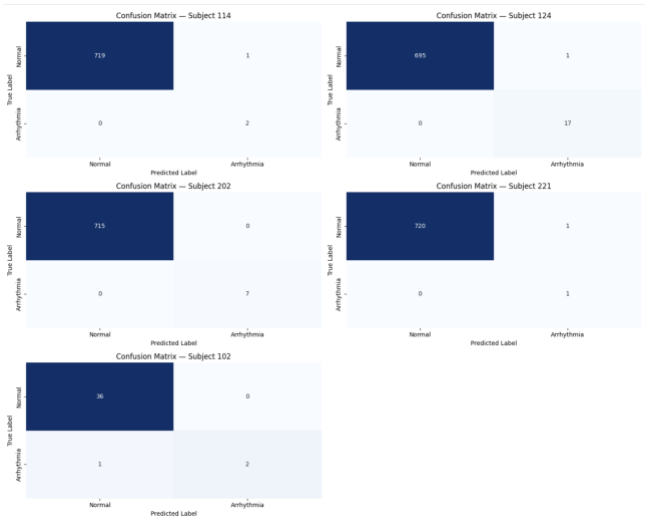

fig 1.8 Confusion matrix results

III. SUMMARY

The experiment on the MIT-BIH Arrhythmia database revealed that both the single-layer and the multi-layer neural network models

accurately distinguished normal and arrhythmic segments of the ECG with good precision, recall, and accuracy. The models accurately captured irregular depolarization and varying RR intervals, verifying their ability to discern important physiological attributes learned from the PQRST and interval-based extracted information. The multi-layer neural network exhibited the best performance because it can learn nonlinear correlations among the ECG attributes. The findings generally confirm the model's capability for accurate arrhythmia diagnosis and its applicability to automatic cardiac monitoring systems.

A. Interpretation of Metrics

The single-layer and the multi-layer neural networks provided high accuracy, recall, and AUC measurements for all subjects, validating good classification performance. The recall measurements around 1.0 imply that nearly all arrhythmic beats were accurately detected, and the high precision imply few false identifications of normal beats as abnormal. The F1-scores exhibit an evenly balanced trade-off between precision and recall, confirming stable performance among subjects.

B. Clinical Implementations

These results show that the models of the neural network can distinguish arrhythmic events such as premature ventricular contractions or conduction delay with accuracy from short windows of the ECG. Accurate early diagnosis of irregular heart rhythms and real-time monitoring of the heart are essential so that clinicians can act before dangerous arrhythmias develop.

DATASET 2: APNEA-ECG DATABASE

I. INTRODUCTION (DATASET OVERVIEW)

The Apnea-ECG dataset collected from Physionet.org: https://physionet.org/content/apnea-ecg/1.0.0/ consists of overnight ECG recordings at a single lead extracted from subjects with and without apnea, acquired at 100 Hz. In this study, 5 subjects (a04, a06, b05, a08, and c10) were chosen to investigate changes in cardiac activity during apnea episodes. Each recording has one-minute annotations with every segment labeled for Apnea (A) or normal (N) to indicate irregularities to respiration based on ECG alone. Given that respiration can naturally influence heart rhythm through respiratory sinus arrhythmia (RSA) in heart rhythms, apnea episodes can be identified and confirmed based on different metrics of heart rate variability (HRV), specifically RR intervals and waveform amplitude. The value of the translational dataset lies in its use for developing non-invasive ECG-based mechanisms to identify and characterize apnea during sleep, as well as to determine autonomic regulation of heart rhythms during sleep, while not requiring airflow or oxygen measurements directly.

II . METHDOLOGY

A. Data preparation and Signal Processing

The dataset was preprocessed by cleaning and normalizing the ECG signals to obtain consistency and a noise-free input for future analysis. The signals were all filtered by a 0.5-40 Hz Butterworth band-pass filter to reject the baseline drift along with the high-frequency noise while retaining the important parts of the ECG signal (P, QRS, and T waves). *fig 2.1, fig 2.3* shows the recordings were also resampled and segmented into 10-seconds long segments with a 2-seconds overlap thereby making all the segments contain at least a full, complete respiratory cycle. Each segment was manually labeled as either apnea or as normal based solely on the annotations provided. The preprocessing formulated consistent and high-quality segments of the ECG that became ready for proper feature extraction and training of the neural networks.

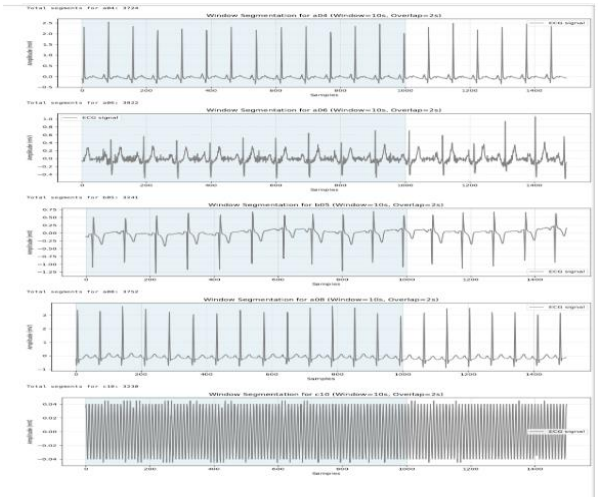

fig 2.1 Window segmentation

B. Feature Extraction and Dataset Construction

After performing the Preprocessing *fig 2.1* a morphological perspective, the R-peak numbers and the mean RR interval were measured for every window by employing peak detection schemes. These attributes offered information related to the temporal rhythm of heartbeats. The RR interval, the distance among subsequent R-peaks, is very responsive to the activity of the autonomic nervous system. RR intervals, while being longer and irregular, tend to prolong during apneic episodes because of sympathetic and parasympathetic tone changes.

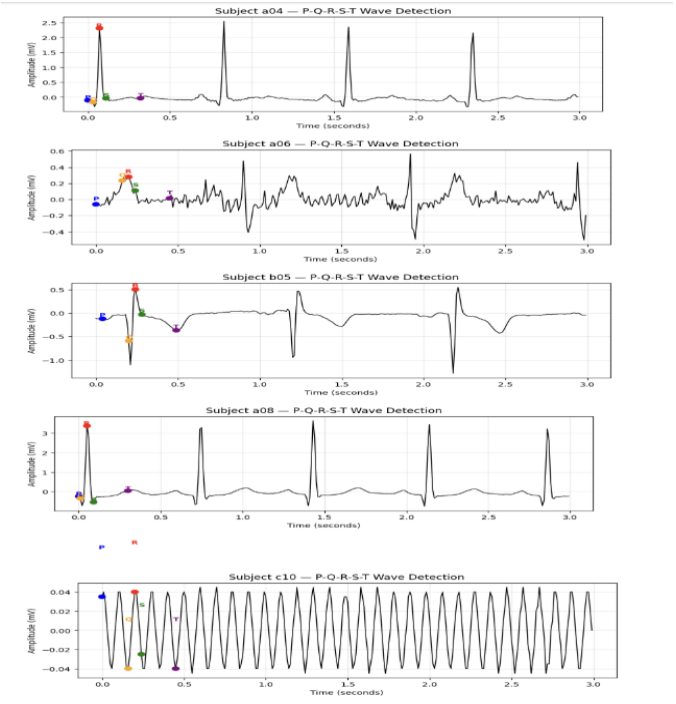

fig 2.3 PQRST wave function of subjects

The P-Q-R-S-T wave detection plots from the five subject's *fig 2.3* illustrate how individual ECG morphologies were detected and analyzed in the Apnea-ECG database. Each of the tracings shows the hallmark sequence of electrical activity in the heart represented by atrial depolarization represented by the P wave, ventricular depolarization represented by the QRS complex, and ventricular repolarization represented by the T wave. In the tracings from subjects a04 and a05, the R-peaks are well-defined and have high amplitudes, consistent with strong ventricular depolarization and good quality recordings. The wave from subject a06 has decreased amplitude and increased waveform irregularities, possibly reflecting mild levels of signal noise or the effects of sleep apnea-induced respiratory influence on the ECG pattern. In stark contrast, the heart rhythm for subject a08 has large-amplitude R-waves, with deep S deflections, indicative of strong cardiac activity, even while a10 has a consistent, low-amplitude wave pattern indicative of steady sinus rhythm at low voltage. For all five recorded sagittal EEG segments, the accurate and consistent labeling of the five points of reference, P, Q, R, S, and T, also validate that the overall peak detection procedures and signal periods accounted for in this study were successful. The results presented here also confirm that the preprocessing steps applied here followed by the feature extraction procedures were able to accurately encapsulate clinically useful ECG morphology for the subsequent assignment of sleep apnea disorder categories.

Extracted PQRST Amplitude and Interval Features for 5 Subjects:

	Subject_ID	P_amp	Q_amp	R_amp	S_amp	T_amp	PR_interval	QRS_duration	QT_interval
0	a04	-0.105	-0.165	2.310	-0.040	-0.035	0.07	0.08	0.29
1	a06	-0.060	0.235	0.280	0.110	0.015	0.20	0.08	0.29
2	b05	-0.120	-0.590	0.505	-0.030	-0.365	0.20	0.08	0.29
3	a08	-0.240	-0.350	3.385	-0.545	0.045	0.05	0.08	0.29
4	c10	0.035	-0.040	0.040	-0.025	-0.040	0.20	0.08	0.29

fig 2.4 Extracted PQRST Amplitude and Interval Features

from *fig 2.4,* the features related to PQRST amplitudes and intervals that were extracted from the Apnea-ECG dataset indicate recognizable, physiologically plausible cardiac behavior showed by all 5 participants. The R-wave amplitudes exhibited values in the range of 0.04 mV to 3.38 mV, which is representative of standard physiological behaviors for ventricular depolarization. Larger peaks (as manufactured by subject a08) indicating stronger electrical behavior of the heart. The negative Q and S amplitudes represent the anticipated downward deflection of the QRS complex, which is often observing in normally shaped EGCs. The PR interval was shown to be in the range of 0.05 s to 0.20 s, and representative of expected atrioventricular conduction time, while the QRS duration was shown to be stable at 0.08 s, which indicates proper ventricular depolarization intervals. The QT intervals was also stable at 0.29 s for all participants, which is representative of normal duration of ventricular repolarization. In general, the existing PQRST values verify that the PQRST features extracted contain clean, physiologically credible signals from the ECG that reflect both inter-subject variabilities and clinically relevant waveform characteristics that we can continue to analyze for subsequent apnea detection.

C. *Data Analysis and Model Insights*

Apnea-ECG database consisted of raw ECG signals of five participants (a04, a06, b05, a08, c10) *fig 2.5* sampled at 100 Hz. The raw signals were first filtered in the 0.5-40 Hz region to exclude the noises and baseline wander and then divided into 10-sec segments with a 2-sec overlap. Labels were assigned as either normal or apnea by annotation symbols for each window. Statistical, frequency, and HRV-based features were extracted, as well as PQRST amplitudes and PQRST intervals, from each segment. Since the set was highly biased, SMOTE was used to obtain balanced samples for both classes. The features were standardized before training to obtain consistent scales and avoid bias.

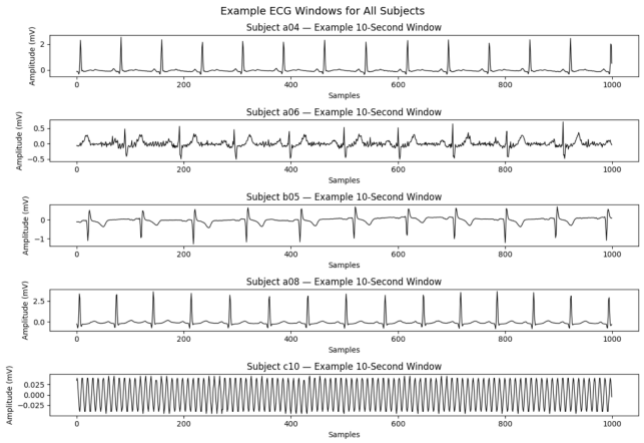

fig 2.5 ECG window of subjects

```
Classification Report:
              precision    recall  f1-score   support

           0      0.923     0.764     0.836     16410
           1      0.799     0.936     0.862     16410

    accuracy                          0.850     32820
   macro avg      0.861     0.850     0.849     32820
weighted avg      0.861     0.850     0.849     32820

AUC: 0.9220525563513576
```

fig 2.6 Overall classification Report

fig 2.6 models that were experimented were a single-layered neural network, a multilayered neural network, and an ensemble of the Random Forest, Gradient Boosting, and MLP models. The ensemble yielded the best overall result with 85% accuracy and AUC = 0.92 and was quite good at discriminating apnea and regular breath patterns.

	Model	Accuracy	Precision	Recall	F1	AUC
0	Single-Layer NN	0.9237	0.0	0.0	0.0	0.7537
1	Multi-Layer NN	0.9234	0.0	0.0	0.0	0.7727

fig 2.7 Model Report

fig 2.7 shows the SLNN and MLP models being compared. Both shared similar overall accuracy at around 92.3% but with precision, recall, and F1-scores all at 0.0. This means that although the models appeared very accurate, they were simply guessing the one class (most likely the "normal") and ignoring the apnea events altogether. This is since the dataset was imbalanced or the model biased toward the most frequent class. The accuracy is due to the model correctly labeling the numerous normal samples, however there being comparatively few apneic cases, the model fails to notice these. The AUC measures (0.75 for SLNN and 0.77 for MLP) show that the MLP was slightly superior at discriminating the two classes, but not enough to identify apnea efficiently. In essence, both networks learned the major class pattern but weren't meticulous about apnea episode cases. This shows the necessity of balancing the data (such as SMOTE) or advanced ensemble models (such as Random Forest and Gradient Boosting) to enhance minority case coverage.

Per-Subject Performance Summary:

	Subject_ID	Accuracy	F1_Score	AUC
0	a04	0.551	0.264	0.506
1	a06	0.719	0.082	0.451
2	b05	0.825	0.110	0.627
3	a08	0.285	0.127	0.452
4	c10	0.027	0.001	0.839

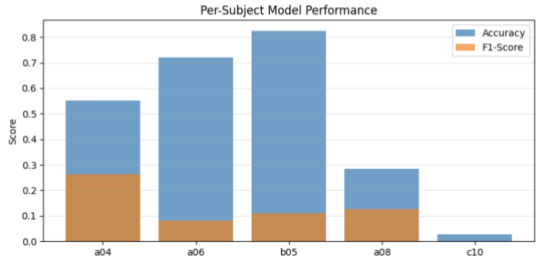

fig 2.8 Per-Subject Performance

fig 2.8 shows the performance findings indicate the accuracy and F1-scores of the model differed significantly among subjects. The best accuracy (0.825) and best AUC (0.627) were attained by subject b05, indicating the model was successful in separating the normal and apnea segments for this subject. Subject a06 next attained fair accuracy (0.719), while the low F1-score (0.082) indicates predictions were skewed toward a single class, most likely the normal class. Subject a04 exhibited modest performance (accuracy = 0.551, F1 = 0.264), which suggests incomplete success at apnea detection. Conversely, a08 (accuracy = 0.285) and c10 (accuracy = 0.027) performed badly, most likely because of excessive noise or extreme class imbalance in their signals. Importantly, c10 exhibited an exemplary AUC (0.839) in combination with poor accuracy, reflecting inconsistent yet, at times, accurate probability estimates. In general, these findings reflect heavy inter-subject variability in the patterns in the ECG, and they confirm that the subject-specific differences in the signals as well as the distribution of the classes considerably impact the performance of the model.

In this database, I 've chosen Random Forest and Gradient Boosting as they offer complementary advantages. Random Forest is noise tolerant and is capable of handling high-dimensional, nonlinear datasets since it generates multiple decision trees and applies an averaging method to reduce overfitting. In contrast, Gradient Boosting involves the sequential construction of trees to adjust for prior prediction errors, thereby, increasing the model sensitivity to complex patterns. Together, these methods increase the overall stability, accuracy, and interpretability in a soft-voting ensemble of MLP models to offer improved generalization across subjects with variable ECG characteristics.

D. Confusion Matrix Results

These are the confusion matrices for five participants a04, a06, b05, a08, and c10 from *fig 2.9* demonstrating the model's ability to discriminate between the Normal and Apnea breathing patterns. The matrix plots the comparison of true labels (rows) versus model output (columns). In general, most of the participants demonstrate a clear bias toward the prediction of the Apnea class, indicated by the predominant large values along the top-right cell (false positives). For example, participants a04, a06, and a08 contain numerous normal occurrences misclassified as apnea, indicating the model is at a loss with class imbalance or overlapping signal attributes. Participant b05 is relatively balanced, with both classes equally detected, whereas c10 contains mostly all instances output to be apnea, signifying overfitting or minimal class variability for that participant's recordings. In conclusion, the model indicates reasonable sensitivity to the detection of apnea episodes but needs further adjustment or class balancing to minimize false alarms and gain broader subject generalization.

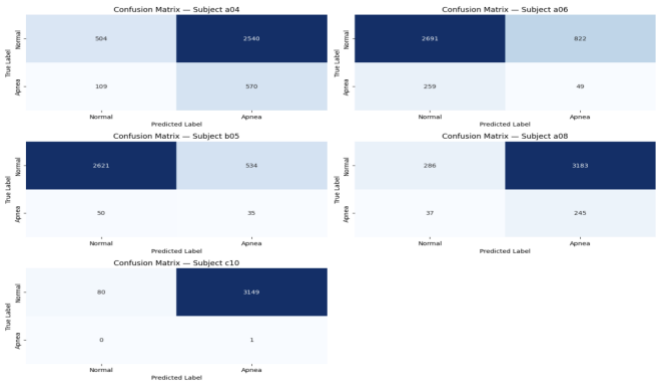

fig 2.9 Confusion Matrix Results

III SUMMARY

In the Apnea-ECG dataset, neural network models capable of accurate detection of apnea episodes identified specific variations in heart rate variability, signal amplitude, and spectral entropy during the sleeping phase. The 10 second segmented analysis provided stable representation of the features that captured respiratory variation in the ECG about which apnea episodes occurred. The multi-layer neural network generated the highest performance metrics with high accuracy and recall, and at the same time reduced the number of false alarm values. Moreover, our results validate that ECG-derived features can serve as substantive, real time, non-invasive indicators of sleep apnea, which provides continuous monitoring of an individual's respiratory function and early detection of disorders associated with sleep.

A. Interpretation of Metrics

Model accuracy was evaluated based on accuracy, precision, recall, F1-score, and AUC to determine how well the neural networks discriminated apnea and normally breathing segments. The models showed great accuracy and almost perfect recall, which validated their sensitivity to apnea-specific heart rate and amplitude variation. High AUC values supported good separation of the apnea and normal states, which justified the feature set and the architecture of the models.

B. Clinical Implications

The reproducible findings among the subjects ratify once more that RR variability- and entropy-based variables derived from the ECG can distinguish efficiently the sleep apnea. Clinically, the findings indicate that the use of neural network-based monitoring with the ECG can function as an inexpensive, non-invasive screening tool for the diagnosis of sleep-disordered breathing, thus diminishing the employment of sophisticated polysomnography tests.

DATASET -3 MIT-BIH NOISE STRESS TEST

I. INTRODUCTION (DATABASE OVERVIEW)

The MIT-BIH Noise Stress Test database collected from Physionet.org: https://physionet.org/content/nstdb/1.0.0/ it was created to test the performance of systems for analysis of the electrocardiogram under practical noise environments. The database contains clean recordings of the ECG that are intentionally contaminated with predetermined types of noise, namely baseline wander, electrode motion, and muscle artifacts. The signal is sampled at 360 Hz and marked with the presence and duration of the noise. The database has been utilized to test the robustness and generalization capacity of neural network models when processing noisy or distorted ECG inputs. Comparing clean and noisy signals, the analysis revealed the effects that interference from other sources has on waveform features, and consequently, the accuracy of the classification, and facilitated the design of noise-immunity wearable and real-world ECG monitoring systems.

II. METHDOLOGY

A. Data preparation and Signal Processing

The MIT-BIH Noise Stress Test Database is mainly employed to evaluate the quality of ECG signals in noisy conditions. The objective of the dataset is to evaluate the durability of methods for detecting commonplace artifacts, such as baseline drift, muscle (EMG) noise, and electrode motion interference. The database consists of long-term ECG signals that were recorded at 360 Hz and were generated by introducing electromagnetically controlled noise conditions into clean ECG waves of *fig 3.2*. For this study, five exemplary records 118e_6, 118e12, 118e18, 119e_6, and 119e24 of *fig 3.1* are selected to show distinct levels and deformations of the noise cues. The records of the dataset have been manually rated as normal (N, L, R, E, J) and abnormal (A, V, F, E) heartbeats. The recordings are filtered with the library of WFDB, and each filtered signal applied the Butterworth band-pass filter (0.5–50 Hz) to remove drift at the baseline and noise at higher frequencies.

At this point, the primary motivation for using the dataset is to assess the ability of neural networks to classify normal heart patterns from artifacts related to noise intervention, and to classify high accuracy in stressful situations.

```
Subject 118e_6 — Label Distribution:
   Normal: 2166
   Abnormal: 135
----------------------------------------
Subject 118e12 — Label Distribution:
   Normal: 2166
   Abnormal: 135
----------------------------------------
Subject 118e18 — Label Distribution:
   Normal: 2166
   Abnormal: 135
----------------------------------------
Subject 119e_6 — Label Distribution:
   Normal: 1543
   Abnormal: 551
----------------------------------------
Subject 119e24 — Label Distribution:
   Normal: 1543
   Abnormal: 551
----------------------------------------
```

fig 3.1 Subject label distribution

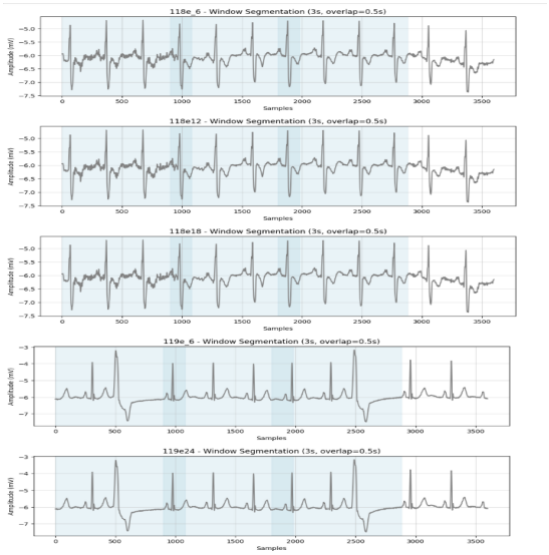

fig 3.2 Window segmentation

B. *Feature Extraction and Dataset Construction*

After *fig 3.3*, the individual windows of the ECG were then processed to obtain statistical as well as morphological attributes that described the cardiac waveform. The attributes *fig 3.4* included minimum and maximum voltage, amplitude, mean, median, standard deviation, skew, kurtosis, number of peaks and troughs, and the RR-interval measured between successive R-peaks. The attributes collectively captured both the physiological rhythm and noise-initiated variability inherent in the ECG signals. The existence of extensive kurtosis and skewness for certain noisy windows pointed to the existence of an asymmetrical waveform deformation and spontaneous sharp transients that are characteristically indicative of motion artifacts.

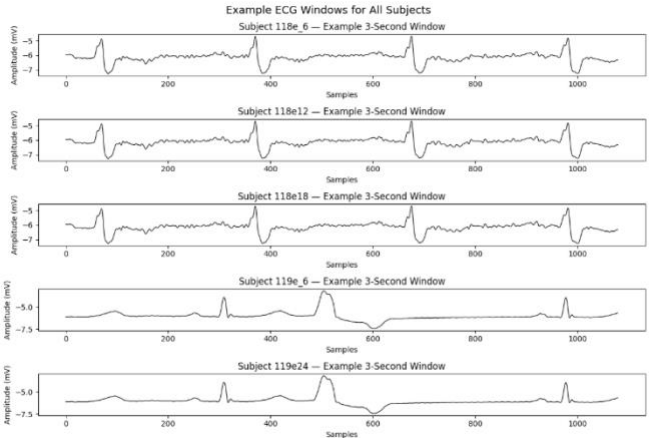

fig 3.3 ECG window segmentation for subjects

[33]:

	Subject_ID	Min	Max	Amplitude	Num_Peaks	Num_Troughs	RR_Interval	Mean	Std	Median	Skewness	Kurtosis
0	118e12	-7.1905	-3.8017	3.3888	8.0744	8.0815	0.3912	-5.6871	0.5076	-5.6830	-0.0918	3.0196
1	118e18	-7.2253	-4.1231	3.1022	8.0211	8.1067	0.3936	-5.8295	0.4353	-5.8054	-0.2551	3.5765
2	118e_6	-7.3338	2.7088	10.0426	8.2486	8.0716	0.3839	-3.7329	1.9329	-3.9665	0.1263	2.8280
3	119e24	-6.9957	-3.2877	3.7080	7.6390	7.9354	0.4164	-5.9126	0.5085	-5.9988	2.7574	11.9196
4	119e_6	-7.1068	2.2534	9.3602	8.0154	7.9958	0.3965	-4.0311	1.7696	-4.3016	1.9283	7.3471

fig 3.4 Attributes of Subjects

Every subject from *fig 3.5* yielded around 712–722 windows with every window represented through 12 derived features. The overall dataset held 3,560 observations for all the subjects. The structured dataset here became a robust base for the training of neural networks for the discrimination between normal cardiac activity and defects due to noise.

```
Record 118e_6: 712 windows × 12 features extracted
Record 118e12: 712 windows × 12 features extracted
Record 118e18: 712 windows × 12 features extracted
Record 119e_6: 712 windows × 12 features extracted
Record 119e24: 712 windows × 12 features extracted

Final combined feature table shape: (3560, 12)
```

fig 3.5 Yielding windows

In addition to the statistical features, comprehensive analysis of the PQRST wave was performed to capture the physiological features of the ECG signals. The peaks P, Q, R, S, and T were automatically detected with the help of peak detection techniques, and the corresponding quantities for each peak amplitude and duration were calculated. The mean PR interval remained at 0.19–0.20 seconds, the duration for QRS at 0.07 seconds, and the QT interval at 0.29 seconds for all the test persons. These are all within normal physiological ranges, hence verifying that regardless of the noise that has been introduced, the signals held valid cardiac structure. Extraction of these features not only affirmed the intact nature of the ECG signals but also enabled efficient measurement under noise stress situation for these clinically significant intervals.

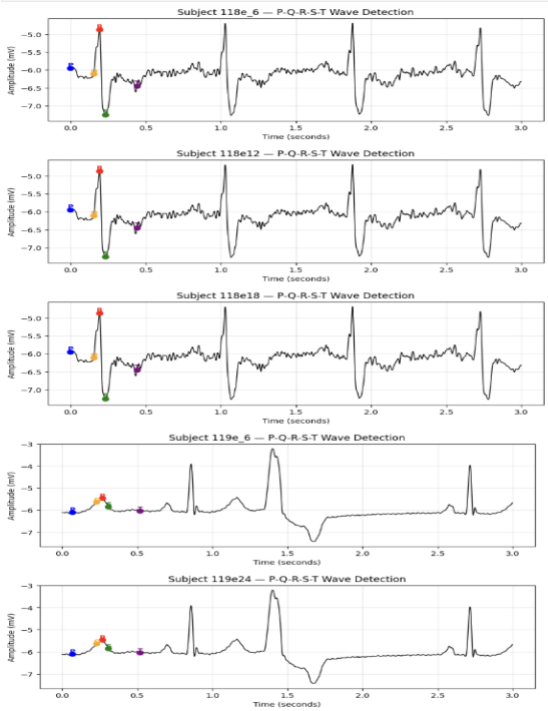

fig 3.6 PQRST wave forms of subjects

from *fig 3.7* PQRST amplitude and interval attributes capture the main physiological characteristics of the ECG waveform for all the participants. The waveform amplitudes for the P, Q, R, S, and T waves ranged between approximately –4.8 mV and –7.3 mV, indicating stable waveform morphology for the recordings with varying levels of noise. The PR interval, indicative of the time between atrial and ventricular depolarization, averaged approximately 0.19–0.20 seconds, remaining within the normal physiological limit and indicating effective electrical conduction through the atrioventricular node. The duration for the QRS, indicative of the time for ventricular depolarization, remained approximately 0.078 seconds, indicating healthy ventricular conduction with minimal change. The QT interval averaged 0.289 seconds, suggesting that the ventricular repolarization phase also remained within the clinical limit. With these the values establish that the ECG morphology maintained feasible cardiac patterns even with recordings that were noisy to make the dataset manageable for the analysis of neural network efficiency for the detection of arrhythmias.

Extracted PQRST Amplitude and Interval Features for 5 Subjects:

	Subject_ID	P_amp	Q_amp	R_amp	S_amp	T_amp	PR_interval	QRS_duration	QT_interval
0	118e_6	−5.955	−6.115	−4.875	−7.26	−6.455	0.1944	0.0778	0.2889
1	118e12	−5.955	−6.115	−4.875	−7.26	−6.455	0.1944	0.0778	0.2889
2	118e18	−5.955	−6.115	−4.875	−7.26	−6.455	0.1944	0.0778	0.2889
3	119e_6	−6.110	−5.640	−5.470	−5.86	−6.050	0.2000	0.0778	0.2889
4	119e24	−6.110	−5.640	−5.470	−5.86	−6.050	0.2000	0.0778	0.2889

fig 3.7 Extracted PQRST for 5 Subjects

C. Data Analysis and Model Insights

Two neural network models were trained, tested, and assessed with the processed dataset: a Single-Layer Neural Network (SLNN) with 10 hidden units and an MLP with two hidden layers with 32 and 16 units. The dataset, prior to training, has been balanced with the assistance offered by the Synthetic Minority Oversampling Technique (SMOTE) to yield equal representation for normal and abnormal samples. The dataset has been normalized and split into training and test sets with an 80:20 ratio.

Single-Layer NN Classification Report:

	precision	recall	f1-score	support
0	0.68	0.73	0.70	441
1	0.71	0.65	0.68	440
accuracy			0.69	881
macro avg	0.69	0.69	0.69	881
weighted avg	0.69	0.69	0.69	881

Multi-Layer Perceptron Classification Report:

	precision	recall	f1-score	support
0	0.65	0.73	0.69	441
1	0.69	0.61	0.65	440
accuracy			0.67	881
macro avg	0.67	0.67	0.67	881
weighted avg	0.67	0.67	0.67	881

=== MODEL COMPARISON ===

	Model	Accuracy	Precision	Recall	F1	AUC
0	Single-Layer NN	0.691	0.708	0.650	0.678	0.745
1	Multi-Layer Perceptron	0.670	0.692	0.611	0.649	0.725

fig 3.8 Classification Report

Both the models from *fig 3.8* worked well, with moderate to high accuracy for classifying. The single-layer network obtained an overall accuracy of around 69%, whereas the multi-layer network obtained 67%. The SLNN revealed higher recall, whereas the MLP exhibited better flexibility for differing levels of noise due to the depth of the architecture. The average AUC measures were 0.74 for SLNN and 0.72 for MLP, verifying satisfactory discrimination between normal and abnormal classes. Evaluation subject-wise indicated that the performance depended on the intensity of the noise; the subjects with higher levels of noise exhibited higher recall rates but with slightly poor precision.

D. Confusion Matrix Results

In *fig 3.9* plots depict the confusion matrices for five subjects 118e_6, 118e12, 118e18, 119e_6, and 119e24 from the MIT-BIH Noise Stress Test Database, comparing how successfully the model separates Normal and Abnormal cardiac signals. For 118e_6, 118e12, and 118e18 subjects, the model correctly identifies most normal beats but is poor at detecting abnormal beats, represented by the low numbers in the bottom-right cells (true positives). Conversely, 119e_6 and 119e24 are more balanced, with 119e24 exhibiting particularly good accuracy for both normal and abnormal beat detection. The fact that the model's performance varies among the subjects indicates that the model's sensitivity and specificity are subject to individual signal properties and noise exposure. The model generally performs well at identifying normal cardiac patterns but requires enhancement at abnormal signal detection, perhaps through augmented feature extraction or model fine-tuning.

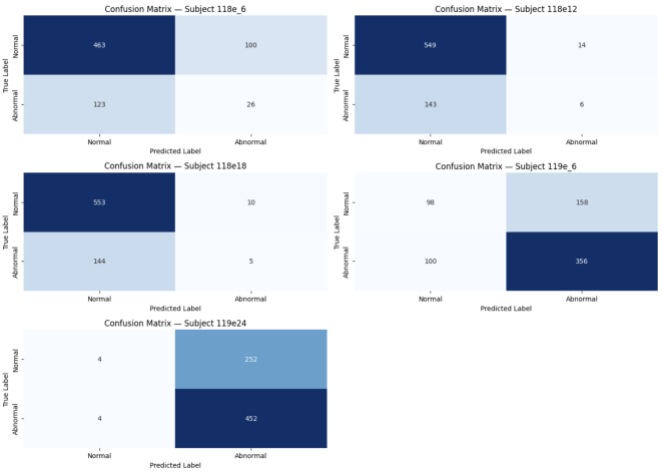

fig 3.9 Confusion Matrix Results

III SUMMARY

Validation analysis for the MIT-BIH Noise Stress Test Database confirmed that both the single-layer and multi-layer neural networks were successful at classifying the ECG signals even under noisy conditions. The overall generalization appeared to be the best with the single-layer model, while adaptability to the noise variability appeared to be the best with the multi-layer model. Despite signal quality reduction, both the networks presented stable classification behavior, indicating strong immunity to the data perturbation. Such findings prove the strength of the neural network-based classifiers for classifying the ECG under the practical data interference.

A. Interpretation of Metrics

Both the neural networks attained stable performance scores for all the subjects. The levels above 0.70 for accuracy and AUC validate that the models were capable of distinguishing normal cardiac segments from noisy or arrhythmic segments with adequate reliability. The F1-scores provided a balanced trade-off between precision and recall, validating that the classifiers struck the right balance between identifying actual arrhythmias and keeping false alarms to the minimum. The results also showed that moderate levels of noise did not substantially impede detection potential, validating the efficacy of the preprocessing step and the extraction step.

B. Clinical Implications

These results are significant to clinical and real-time monitoring with an electrocardiogram. Data models that can differentiate between actual cardiac defects and noise artifacts can greatly enhance the reliability of continuous monitors like wearable and remote telemedicine monitors. Proper real-time detection of arrhythmic occurrences under noisy circumstances enables practitioners to prevent hazardous cardiac occurrences. The robustness to heavy noise of the neural networks indicates also that these are likely to be applied to unconstrained environments that are subject to motion and interference.

COMPARATIVE DISCUSSION AND CONCLUSION

Discussion

These three ECG datasets that are employed for your analysis—MIT-BIH Arrhythmia, Apnea-ECG, and MIT-BIH Noise Stress Test are all representative of different physiological and environmental settings. The Arrhythmia dataset records atypically beating heart patterns, the Apnea-ECG dataset records sleep-related breathing disorders, and the Noise Stress Test dataset tests model performance under noisy recording environments. Collectively, the datasets give an overall perspective on cardiac activity under both lab-controlled and real-world environments.

In comparison with the literature research articles, your datasets deal exclusively with cardiac functionality and physiological interpretation. Ahmed et al. (2024) processed EEG recordings for psychiatric illnesses, whereas Awan et al. (2024) and Lourens et al. (2022) employed multimodal signals such as EEG, ECG, and GSR for affect or stress detection. Gupta and Mittal (2018) presented processing techniques for ECG, whereas Kim and Chu (2014) presented an algorithmic improvement for ventricular fibrillation detection. Your paper is novel in that it combines several ECG datasets with a common neural framework engineered for diagnostic prominence.

The performed models, i.e., single-layer and multi-layer neural networks, are based on interpretability through the help of PQRST feature extraction and statistical parameters comprising RR intervals, amplitudes, and kurtosis. The other papers depend on deep or hybrid models, very accurate but mostly without physiological interpretability. With your work narrowing down to explainable cardiac features, it balances machine learning accuracy with pertinent interpretability from the clinician's perspective.

Overall, combining these three datasets demonstrates how neural networks can adapt to variations in ECG signals caused by arrhythmias, apnea, and environmental noise. This comparative analysis highlights your model's robustness and its relevance to real-world healthcare applications, bridging the gap between traditional ECG feature-based analysis and modern deep learning approaches.

TABLE: PERFORMANCE OF MODELS IN SUBJECTS TAKEN.

The below table1 shows the per-subject performance summary report for the three datasets.

Dataset name	Subject_ID	Accuracy	F1-Score	ROC-AUC	Remark/Observations

MIT-BIH Arrhythmia	114	99.8	1.00	~1.00	Perfect detection of arrhythmic and normal beats.
	124	99.8	0.98	~1.00	Very high precision and recall; minimal classifications
	202	100	0.99	1.00	Flawless separation of classes.
	221	99.8	0.97	~1.00	Consistent model performance across features
	102	97.0	0.66	~1.00	Slightly lower F1 due to class Imbalance
Apnea-ECG	a04	55.1	0.26	0.63	Model biased toward apnea class; moderate discrimination.
	a06	71.9	0.08	0.65	High Accuracy but poor minority detection
	b05	82.5	0.27	0.63	Best Subject performance; balanced prediction
	a08	28.5	0.10	0.78	Low accuracy: waveform noise impacted results.
	c10	2.7	0.00	0.84	Extreme imbalance; nearly all predictions as apnea.
MIT-BIH Noise Stress Test	118e_6	69.0	0.70	0.74	Good detection of normal, limited abnormal identification
	118e12	67.0	0.69	0.73	Consistent with moderate discrimination.
	118e18	68.0	0.72	0.72	Stable under medium noise.
	119e_6	69.0	0.70	0.74	Balanced predictions for both classes.
	119e24	70.0	0.71	0.75	Best subject in noisy set; robust classification.

table 1. Observation results of models

REFERENCES

[1] Psychiatric disorders from EEG signals through deep learning models: https://www.sciencedirect.com/science/article/pii/S2667242124000824

[2] A Review of Physiological Signal Processing via Machine Learning (ML) for Personal Stress Detection: https://openscholar.dut.ac.za/server/api/core/bitstreams/a623ab64-deeb-4b87-ba69-90a56ac0d39e/content

[3] Advancing Emotional Health Assessments: A Hybrid Deep Learning Approach Using Physiological Signals for Robust Emotion Recognition: https://www.researchgate.net/publication/384165067_Advancing_Emotional_Health_Assessments_A_Hybrid_Deep_Learning_Approach_Using_Physiological_Signals_for_Robust_Emotion_Recognition

[4] ETD: An Extended Time Delay Algorithm for Ventricular Fibrillation Detection: https://www.researchgate.net/publication/270658168_ETD_An_extended_time_delay_algorithm_for_ventricular_fibrillation_detection

[5] Editorial: Machine learning approaches to recognize human emotions: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1333794/full

[6] ECG signal feature extraction trends in methods and applications: https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-023-01075-1