Adolfo, Cid Mathew, Chizari, Hassan ORCID logoORCID: https://orcid.org/0000-0002-6253-1822, Win, Thu Yein ORCID logoORCID: https://orcid.org/0000-0002-4977-0511 and Al-Majeed, Salah ORCID logoORCID: https://orcid.org/0000-0002-5932-9658 (2021) Sample Reduction for Physiological Data Analysis Using Principal Component Analysis in Artificial Neural Network. Applied Sciences, 11 (17). Art 8240. doi:10.3390/app11178240

Sample Reduction for Physiological Data Analysis Using Principal Component Analysis in Artificial Neural Network

Cid Mathew Santiago Adolfo, Hassan Chizari, Thu Yein Win and Salah Al-Majeed

Abstract: With its potential, extensive data analysis is a vital part of biomedical applications and of medical practitioner interpretations, as data analysis ensures the integrity of multidimensional datasets and improves classification accuracy; however, with machine learning, the integrity of the sources is compromised when the acquired data pose a significant threat in diagnosing and analysing such information, such as by including noisy and biased samples in the multidimensional datasets. Removing noisy samples in dirty datasets is integral to and crucial in biomedical applications, such as the classification and prediction problems using artificial neural networks (ANNs) in the body's physiological signal analysis. In this study, we developed a methodology to identify and remove noisy data from a dataset before addressing the classification problem of an artificial neural network (ANN) by proposing the use of the principal component analysis–sample reduction process (PCA–SRP) to improve its performance as a datacleaning agent. We first discuss the theoretical background to this data-cleansing methodology in the classification problem of an artificial neural network (ANN). Then, we discuss how the PCA is used in data-cleansing techniques through a sample reduction process (SRP) using various publicly available biomedical datasets with different samples and feature sizes. Lastly, the cleaned datasets were tested through the following: PCA–SRP in ANN accuracy comparison testing, sensitivity vs. specificity testing, receiver operating characteristic (ROC) curve testing, and accuracy vs. additional random sample testing. The results show a significant improvement in the classification of ANNs using the developed methodology and suggested a recommended range of selectivity (Sc) factors for typical cleaning and ANN applications. Our approach successfully cleaned the noisy biomedical multidimensional datasets and yielded up to an 8% increase in accuracy with the aid of the Python language.

Keywords: principal component analysis (PCA); artificial neural network (ANN); multidimensional dataset; dimension reduction process; sample reduction process (SRP); receiver operating characteristic (ROC) curve; selectivity (Sc); sensitivity; specificity

1. Introduction

It is clear that biomedical sensors will keep improving exponentially, as they become more accessible; more readily available on the market; more intelligent; smaller and more compact; and integrated into personal belongings, such as cellular phones, watches, and eyeglasses. It is an outstanding contribution to human health, entertainment, the military, security, sports, and leisure and in analysing a patient's physiological data and their interpretation. These sensors are an integral part of biomedical devices; however, this innovation has an ever-evolving trend coupled along with challenges to perform intelligently [1]. The integration of these sensors, subjected to different environments in both wired and wireless implementations, adds random noise to the system. The sudden upswing of the noise floor yields a degradation in the signal-to-noise ratio (C/N) and low energy bit per noise reference (Eb/No) during data processing. The drastic increase in the probability of error (Pe) eventually decreases the accuracy in such a system [2]. This interference results in flawed calculation and decision making, especially for artificial neural networks (ANNs) in biomedical applications [3].

An artificial neural network (ANN) is a part of a computing system that simulates the ability of the human neuron to learn the complex characteristics of the environment, to recognise patterns, and to generalise the interrelationships between the features, including multidimensional datasets. It is part of the umbrella of artificial intelligence (AI) and

deep learning, and it solves problems that would be impossible or difficult to solve by human effort or statistical criteria. ANNs have self-learning skills, enhancing their efficiency as more effective and reliable data become available to them. Primarily, ANNs enable the complex inter-relationships between the features within a given dataset to be identified and have seen widespread adoption in different applications, including biomedical and signal processing, which is readily and publicly available by wearable sensors [4].

Since data are everywhere—such as on the Internet—neglecting the integrity of a source of information poses a significant threat for ANNs to misinterpret the data. For critical aspects of the medical field, the efficacy of a diagnosis is threatened if the analysis that aids medical experts in research has a weak source, especially if dealing with the multitude of datasets that result in the presence of noisy and biased data, which often occurs from seemingly reliable sources [5].

Low-powered sensors are a critical source of unreliable information. Sudden changes in conditions can introduce environmental noise, and there are many possible avenues for noise and interference to blend into any part of the system. A sensor's scalability is another problem, given that the sensors nowadays are low power and with limited computation. Placement or location is another significant constraint, and if they are placed too close to other devices, they may be affected by crosstalk interference [6].

To understand the data, first, one needs to understand their composition. Data comprise valid and dirty data. Accurate data contain information that is accurate, holds predictive power, and is generalised to the entire set. Dirty data contain information that is misleading, noisy, or erroneous data, such as pragmatic contexts and the semantic- and syntactic-biased errors shown in Figure 1 [7].
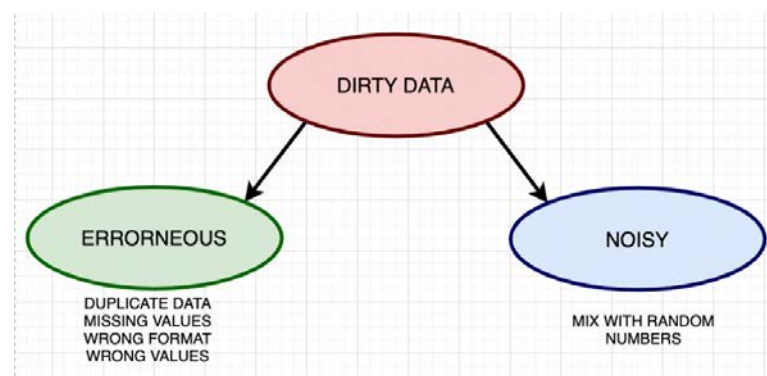
Figure 1. Dirty data.

Noisy data are any random fluctuations considered unwanted, unpleasant, loud, or disruptive that hinder the generalisation of the entire dataset. For an ANN with predictive power to analyse the data, the dataset needs to be free from noise to reduce errors in classification and prediction, which would significantly impact medical experts in their professional biomedical interpretation. Given recent technological solutions, one relevant control algorithm is highlighted in solving these challenges [8].

Principal component analysis (PCA) is usually part of data visualisation. PCA con-verts the features of datasets from high dimensional data and can help convert data to low-dimensional data with the aid of covariance, eigenvalues, and eigenvectors in identifying and sorting the strength of the predictive power of all features [9]. PCA allows for the extraction of critical feature vectors from multidimensional datasets in terms of data visualisation and loses some crucial information along the process [10], which is a considerable disadvantage, primarily if it is used extensively in

the field of machine learning. This study used PCA not as a dimension reduction but as a sample reduction in order to remove the unwanted noisy samples in multidimensional datasets.

This paper proposes an application of a PCA–sample reduction process (SRP) to improve the prediction accuracy of ANNs. Using publicly available biomedical datasets from a different field provides qualitative analysis and demonstrates the effectiveness of said method in improving the accuracy of ANNs. By cleaning the publicly available biomedical datasets before training and testing the ANN, the classification problem is expected to increase the accuracy and to lower the computational cost, which is a great help in dealing with the analysis and prediction of big multidimensional datasets.

A study to reinforce the methodology of an ANN using PCA–sample reduction is proposed in this paper to determine the proposed system's accuracy and performance with different sizes of multidimensional biomedical datasets. Furthermore, we investigate heart disease, voice and speech analysis for gender recognition, breast cancer classification, and cancer patient datasets to prove the versatility and flexibility of the proposed data-cleansing technique.

The rest of the paper is structured as follows. Section 2 provides a literature review of data cleaning in biomedical applications. Section 3 discusses the basic concepts of principal component analysis and related topics. Section 4 discusses how the PCA–SRP is integrated with an ANN. Section 5 provides the recommended Sc ranges. Section 6 interprets the result of publicly available biomedical datasets. Lastly, Section 7 provides a discussion, the conclusion, and future research directions.

2. Data-Cleaning Applications

The volume of data collected nowadays is vastly increasing, and since most data acquired are polluted, the dependability of the data is declining. Various data-cleaning methodologies are available to rectify this issue, but data cleansing remains difficult when working with large data requirements. Data cleaning, also known as data cleansing, is no longer a recent area of research. It aims to increase data quality by detecting and eliminating errors and inconsistencies [11]. As of now, there are two classifications of data cleansing: traditional data cleansing and data cleansing for big data. Traditional data cleansing techniques are so called because it is not used to manage massive volumes of data, such as Potter's Wheel and Intelliclean [12].

Meanwhile, the techniques in Table 1 such as Cleanix [13], SCARE [14], KATARA [15], and BigDansing [16] are developed specifically for big data. Regarding the emerging trends in data-cleaning techniques, one of the new challenges that researchers are about to face is scalability [17]. One of the perennial problems in data analytics is identifying and restoring dirty data, and failure to do so will result in faulty analytics and unreliable decisions. New abstractions and scalability are among the various facets of this issue and are considered when developing data cleaning methods to cope with the amount and diversity of data [13,14,16]—see Table 2. Given the significant amount of data, it needs time to be processed to be suitable for big data analysis and decision making. The data's volume, veracity, and velocity must also be considered when analysing the proposed approaches; however, the researchers mentioned that "Data analytics is not about having the information known, but about discovering the predictive power behind the data collected" [12].

Table 1. Data-cleansing methods for big data.

Methods	Key Features	Approach
Cleanix	Scalability, unification, and Usability	Rule Selection
SCARE	Scalability	Machine Learning Technique
KATARA	Easy Specification, Pattern Validation Data Annotation	Knowledge-base and Crowdsourcing
BigDansing	Efficiency, Scalability and ease of use	Rule Specification

Cleanix, SCARE, and BigDansing focus on the scalability issue in the data cleansing process. Moreover, SCARE and BigDansing do not require any human-domain expert in the cleansing process. SCARE needs an extensive set of rules to update the dataset; however, no expert is present in the process. Nevertheless, the process is expensive, and if the authorities fail to identify correct fixes for the dirty dataset, it will result in redundancy of the training data and a threshold machine learning parameter that is hard to set precisely [15]. Furthermore, BigDansing also requires a set of data-quality rules for optimisation of the cleansing process that requires too many regulations to calibrate and put into place before the start of the cleaning process, as shown in Table 2 below; however, it needs no human-domain expertise to monitor the whole process, although adjusting such parameters is crucial in maintaining the essence of the information in the datasets [12].

Table 2. Currently used data-cleaning technique.

	Volume	Veracity	Velocity
Cleaning Technique Data	**Scalability**	**No need for Extensive Data Quality Rule Optimisation**	**No need for Human Domain Expert**
Cleanix	x		
SCARE	x		x
KATARA		x	x
BigDansing	x		x

These data-cleansing techniques support various data-cleaning tasks such as abnormal value detection and correction, incomplete data filling, de-duplication and conflict resolution (Cleanix), value modification (SCARED), the identification of correct and in-correct data, and the generation of top-k possible repairs for inaccurate data (KATARA) and rules into a series of transformations that enable distributed computations and several optimisations (BigDansing).

There are two kinds of dirty data: erroneous and noisy—see Figure 1. The current big-data-cleansing techniques— Cleanix [13], SCARE [14], KATARA [15], and BigDansing [16]—are focused mostly on erroneous dirty data such as a solution in duplicate entries, missing values, wrong values, and wrong formats. In medical applications, these big-

datacleansing techniques play a critical role in medical record management in medical facilities; however, they do not address noisy data acquired via biomedical sensors. Cleanix, KATARA, and BigDansing are not able to predict the correct values through a machine learning approach and cannot determine how to eliminate the mixed random number [12]; however, the SCARE technique, though constructed via machine learning, can only replace the missing values with the most precise value. When dealing with noisy data, the common approach in applications with machine learning and artificial neural network (ANN) is to address noise reduction and suppression [11,18,19].

Given all of the above data-cleaning techniques, noise and random sample reduction is a significant part of their scope of discussion, which is vital in biomedical data signal analyses acquired in wearable biomedical sensors due to its subjectiveness in a different kind of environment—an environment often filled with varying sorts of noises, such as thermal and acoustic noises and interference. Moreover, since most wearable biomedical sensors are of low power, by suppressing these kinds of noises through signal processing by applying the filtering threshold method, the unsupervised classification is not effective under a low SNR. When the spectral characteristics of the noise are so similar or near that of the sensor-received signal, the detection performance may be degraded [20]. When it applies to the classification problem of an artificial neural network (ANN), obtaining the correct values through comprehensive and extensive quantisation in data-signal processing is essential. Still, it is not sufficient to identify which of the gathered data have a predictive power, as some authors mention that "Data analytics is not about having the information known, but about discovering the predictive power behind the data collected" [12]. Correct values of the data do not guarantee that it holds a valid value in predicting and summarising the entirety of the multidimensional datasets. Mining those correct values is an integral part of data mining in every machine learning and big data analysis.

Complicated and straightforward mistakes are unavoidably present in data input and acquisition. Although much effort could be expended into this front-end procedure to reduce entry mistakes, the truth remains that mistakes in massive datasets are prevalent. Field error rates usually are about 5% or higher unless an organisation takes extraordinary precautions to prevent data inaccuracies [21]. Moreover, this rate is still quite high, at such a rate that it might lead to erroneous interpretation and decision making.

In the case of cleaning the noisy data via biomedical sensors, most researchers use principal component analysis, such as a reduction in dimensionality or feature space [22–26], feature extraction in further data visualisation [27–29], and feature selection tools in ma-chine and deep learning applications [30–34] for machine learning applications.

Principal component analysis (PCA) is usually utilized in dimension or feature reduction and provides a significant increase in accuracy and efficiency along with other machine learning techniques in many applications aside from biomedical application [35–39]. Nevertheless, some of the information is lost during the process of dimension reduction [10].

Figure 2 shows a sample Excel dataset, and PCA reduces the dimensions vertically (by column), shown in red; for the proposed methodology, it reduces the noisy samples horizontally (by rows). We emphasise that the extraction of features, such as covariance, eigenvalues, eigenvectors, and dimension reduction, is not a novel technique that we propose here, but instead, we propose the implementation of sample reduction using PCA. Identifying the noisy samples that cause irregularity in the multidimensional datasets and omitting or reducing them is the main focus of this study.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun
2	0.178491211	0.055165379	0.193881579	0.132171053	0.211842105	0.079671053	2.571840284	12.23283109	0.911291539	0.432742048	0.198486842	0.178491211	0.125876414	0.04701
3	0.21255976	0.064200398	0.233301566	0.19308373	0.26113002	0.06804629	2.632870339	11.52322255	0.910943587	0.506981254	0.266276376	0.21255976	0.108126663	0.01791
4	0.18898524	0.056050744	0.214439024	0.130211382	0.224	0.093788618	3.180794323	14.45188832	0.855932245	0.266642406	0.215804878	0.18898524	0.132015802	0.0472
5	0.178027567	0.070547896	0.19	0.127436897	0.242820513	0.115384615	2.149794566	9.174199803	0.9458361	0.537707592	0.251794872	0.178027567	0.1161125	0.02010
6	0.190421877	0.053285671	0.21543379	0.143196347	0.233333333	0.090136986	2.410619248	9.861366568	0.872305794	0.264242104	0.232054795	0.190421877	0.133339095	0.04785
7	0.165439339	0.058030049	0.181621622	0.103843844	0.199279279	0.095435105	2.45853693	10.32384884	0.909782239	0.46237789	0.184564565	0.165439339	0.103017102	0.0461
8	0.155171185	0.067961896	0.159920213	0.091223404	0.21837766	0.127154255	2.405710157	11.3021319	0.94438662	0.62517986	0.081170213	0.155171185	0.081689643	0.01564
9	0.194216906	0.065705295	0.191339648	0.133748309	0.259539919	0.12579161	2.350557874	9.29922848	0.893323533	0.403060154	0.26105548	0.194216905	0.126158677	0.04780
10	0.193155389	0.066994936	0.209278351	0.138195876	0.252216495	0.114020619	2.613076149	11.09342564	0.906531584	0.462571346	0.140360825	0.193155389	0.143489618	0.05095
11	0.183124768	0.059820084	0.202359882	0.132979351	0.227551622	0.094572271	2.120305466	8.974561921	0.925890098	0.528519618	0.225899705	0.183124768	0.114168261	0.03493
12	0.181555987	0.065929432	0.191240429	0.131209801	0.236834609	0.104624809	2.254150823	9.782985605	0.926799664	0.547278864	0.230689127	0.181555987	0.149405943	0.04692
13	0.141280115	0.074404048	0.13420617	0.090594432	0.202046652	0.11145222	2.052550657	10.94130568	0.966588205	0.744799871	0.060045147	0.141280115	0.109319559	0.02671
14	0.140026485	0.074647541	0.125385253	0.096156758	0.209130075	0.112974316	2.228929212	9.771693746	0.96159251	0.728458548	0.108102734	0.140026485	0.111841991	0.01637
15	0.195680246	0.059588656	0.221933842	0.130025445	0.248296165	0.11826972	2.477426098	9.361352222	0.880763007	0.314705323	0.240814249	0.195680246	0.125406013	0.05429
16	0.182237848	0.076138329	0.206618357	0.138780064	0.232996169	0.094235105	2.696652365	12.9576936	0.912981637	0.448559674	0.043848631	0.182237848	0.139618395	0.01568
17	0.185643489	0.065043383	0.203955432	0.186016713	0.223454039	0.037437326	2.732156446	11.20532255	0.89832992	0.476307945	0.199665738	0.185643489	0.175833394	0.04824
18	0.183433207	0.049426678	0.173185976	0.151097561	0.21277439	0.061676829	2.549621332	12.71592197	0.90747976	0.306336797	0.154085366	0.183433207	0.146007328	0.01530
19	0.181542883	0.059719631	0.190337079	0.125449483	0.215550618	0.09005618	2.391341346	9.381877356	0.898203397	0.396000394	0.186404494	0.181542883	0.127600374	0.04843
20	0.183106973	0.033228357	0.183703704	0.171481481	0.198148148	0.026666667	2.83336536	12.11020215	0.845921349	0.239097676	0.187962963	0.183106973	0.167326447	0.02030
21	0.167761538	0.05888337	0.15993403	0.125607477	0.200879604	0.075272128	2.531500396	11.50948543	0.923640119	0.485117754	0.126838922	0.167761538	0.132089064	0.01853
22	0.166478636	0.06180066	0.181754386	0.108684211	0.202017544	0.093333333	2.481266224	10.28596222	0.916918397	0.519853201	0.166066667	0.166478636	0.113172043	0.0474
23	0.171547961	0.065804842	0.189052632	0.110315789	0.216421053	0.106105263	2.971078153	15.24962235	0.935737882	0.59832555	0.201684211	0.171547961	0.122665736	0.04733
24	0.189954108	0.03720486	0.1921875	0.1765625	0.2065625	0.03	2.744296218	12.26023732	0.850261004	0.264340052	0.1903125	0.189954108	0.182078915	0.0234
25	0.200513095	0.044939774	0.20991832	0.156009335	0.23033839	0.074329905	2.319576234	10.26277699	0.898911684	0.191797697	0.20991832	0.200513095	0.135308185	0.05693
26	0.160925335	0.068626553	0.173488915	0.098833139	0.211225204	0.112392065	2.67033497	12.89559691	0.933988616	0.566987362	0.096872812	0.160925335	0.097513464	0.01990
27	0.213771735	0.041064893	0.22019802	0.203564356	0.236039604	0.032475248	2.911195455	12.76470362	0.845656573	0.279938495	0.222178218	0.213771735	0.183704222	0.04963
28	0.187477037	0.036038199	0.216820513	0.122529915	0.238358974	0.11582926	2.456931539	9.423683865	0.876781326	0.353811879	0.118222222	0.187477037	0.115284365	0.04710
29	0.191055166	0.036377491	0.179898219	0.168142494	0.220152672	0.052010178	3.024759836	13.86349033	0.845371191	0.223577901	0.164223919	0.191055166	0.160947502	0.04757
30	0.203408291	0.038606683	0.209756098	0.194146341	0.224065041	0.029918699	3.027533937	13.23387076	0.850975962	0.272271763	0.204878049	0.203408291	0.164703099	0.04918
31	0.21145984	0.031388715	0.212108434	0.199457831	0.226446783	0.026987952	2.912364805	12.28602411	0.835035487	0.181912572	0.212108434	0.21145984	0.188752092	0.04747

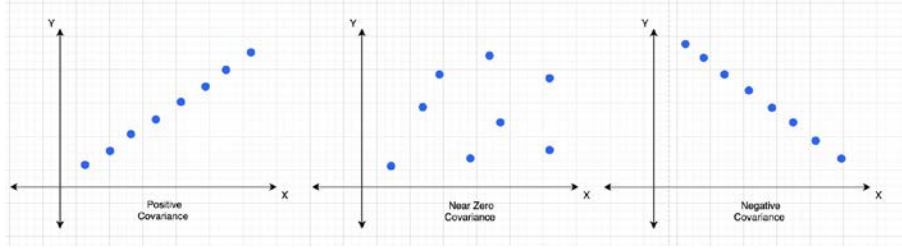**Figure 2.** Sample Excel multidimensional dataset.

The observation and the application of the PCA–sample reduction process are utilised for data cleansing of the noisy multidimensional datasets to increase the accuracy of classification problems in artificial neural networks and to identify recommended threshold ranges. To the best of our knowledge, we are the first to propose this unique approach. We discuss a specific technique for qualitative noise detection and omission of such. These techniques are explained with a motivating example highlighting deep learning classification problems of artificial neural networks in biomedical applications using publicly available biomedical datasets. The simplicity of this technique makes it portable and can apply to a variety of tasks for fast and accurate classification of typical or commonly used artificial neural networks.

In terms of software implementation, developers and data engineers prefer Python programming. It is the best option for projects or programming involving AI and big data analysis as Python is a simple language with a mature and supportive Python community; an abundance of support from renowned corporate sponsors; an extensive and popular selection of libraries; and the ability to work with heavy-hitting frameworks such as TensorFlow, Scikitlearn, OpenCV, and Keras [40].

3. Principal Component Analysis–Sample Reduction Process

Principle component analysis (PCA) aims to minimise the dimensionality of a dataset with many connected variables while keeping as much variance as feasible. The conversion of the new collection of uncorrelated variables known as principal components (PCs) preserves most of the variance included in the original variables. A new set of dimensions or orthogonal measurements are linearly independent and ranked according to the data variance. This means the more crucial principal axis occurs first (more important = more variance/more spread-out data). In general, understanding the PCA, variance, covariance, eigenvalues, and eigenvectors has an essential role in this concept [41].

Figure 3 shows large positive covariance means that X and Y are completely related, i.e., as X increases, Y also increases. Negative covariance portrays the exact opposite relation. Zero covariance means that X and Y have no relation.

Figure 3. Covariance and their values.

The visualisation of data is an excellent approach to understand the patterns that lie in the multidimensional datasets. When information is placed in the horizontal and vertical axes (two-dimensional plane), it is straightforward to understand and discern the pattern that lies in it; however, the difficulty of conceiving it visually in multidimensional data with many features to consider and computing the data analysis becomes complex. Principal component analysis prerequisites require the discovery of patterns between the datasets so that the data are distributed across each dimension by first analysing the contributions of each feature in providing information to the overall dataset through eigenvector analysis. It then performs dimension reduction by keeping the feature columns with the highest eigenvalues.

To demonstrate the efficacy of PCA, Figure 4, which projects a set of multidimensional data onto a two-dimensional space, is used as an example. Due to the high dimensional nature of the data points, it can be challenging to identify a linear correlation between the data points. The points are represented as column vectors before being aggregated into a matrix M. PCA is then applied to M as shown in Equation (1):

$$PCA(M) = W \times V \tag{1}$$

where W and V are the corresponding eigenvalues and eigenvectors, respectively.

V_{PC1} is generated using covariant eigenvalues V with the corresponding top k eigen-value in W, which best represents the data points, and are then selected, before plotting them as shown in Figure 5 [30]. The principal axes (PC_1 and PC_2) are denoted by the red lines passing through the points that provide a graphical representation of the covariance amongst the points.

Lastly, the vital part of the proposed PCA–sample reduction process (PCA_{SRP}) is as follows:

$$PCA_{SRP}(M) = |M \times V_{PC_1}| \tag{2}$$

where $PCA_{SRP}(M)$ is the loading score of each sample in dataset M and V_{PC1} is the PC_1 eigenvectors of the samples in matrix M.

$$S = \sum_{0}^{D} \begin{cases} 1 & \text{if } PCA_{SRP}(M) > Sc \times PCA_{SRP}(M)_{highest} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

S is the number of samples above the loading score of set-biased Sc, and $S \leq D$, where D is the number of samples in the dataset, including the random and noise samples.
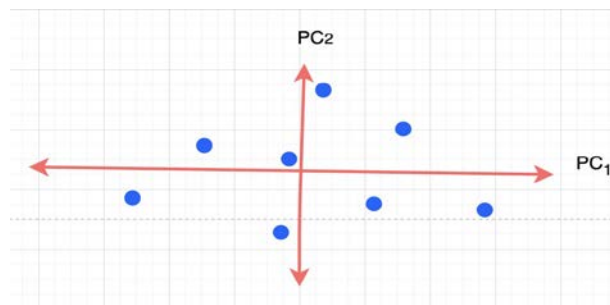
$$passrate = S/D \tag{4}$$

The *passrate* is the rate of samples cleaned using PCA–SRP in dataset M based on Equation (4). Any samples above said threshold are accepted, and those below it will be rejected for the new process dataset D_{PS} as per Equation (3).
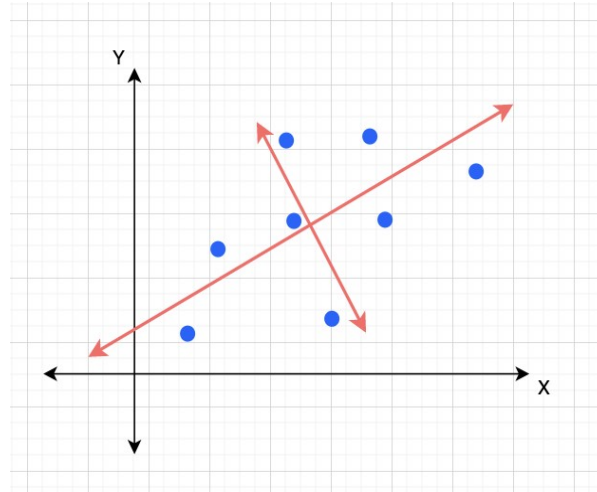
Figure 4. Projection of data points in two dimensions.

Figure 5. Projection of data points in the two newly constructed dimensions.

4. PCA–SRP Implementation in ANN

The proposed framework addresses the limitations of the ANN in processing multidimensional datasets through the use of PCA with the sample reduction process (PCA–SRP). PCA is used to analyse the multidimensional data, and if there exists a significant relation-ship between the features of a dataset, it is arranged from the most significant samples to the least to be visualised and the multidimensional data are put into perspective. It focuses on implementing dimension reduction by removing data from the multidimensional dataset feature columns that do not have high inter-feature covariance; however, the proposed framework using the loading scores can dissect all of the samples or rows in the multidimensional dataset. Figure 6 provides a graphical description of the proposed PCA–SRP based ANN solution.
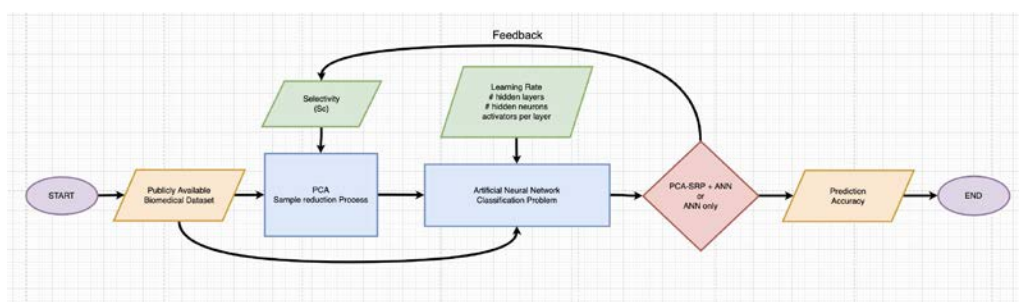
Figure 6. Conceptual flowchart of the proposed approach.

The PCA–SRP Python program generates a new set of multidimensional data with fewer samples based on the screen plot to identify the most significant samples. It is used in the Python ANN program as an input in order to ensure correct classification.

Figures 7 and 8 show the algorithm used in Python in the implementation and analysis of the concept. The implementation starts by extracting the data in .csv format and by injecting it into the PCA–SRP process. Major principal components are extracted through the threshold mentioned above based on the input dataset, resulting in a new set of multidimensional datasets used in ANN.
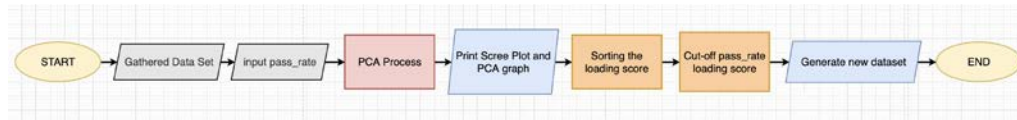
Figure 7. Detailed breakdown of the steps in PCA.
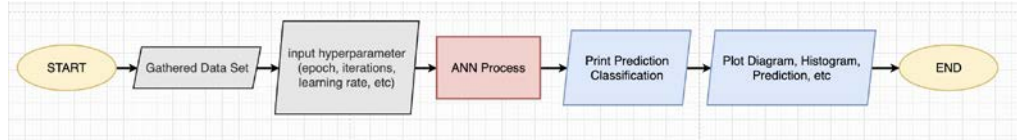
Figure 8. Detailed breakdown of the steps in ANN.

The proposed PCA–SRP utilised in the ANN consists of two hidden layers consisting of 32 neurons and 16 neurons. The learning rate was set at 0.1 and trained for 100 epochs. The input parameters in the PCA–SRP-based ANN are shown in Table 3 [42].

Table 3. PCA–SRP and ANN parameters.

PCA–SRP and ANN Parameters	Values
Most significant sample size	$>Sc \times (PCA_{SRP}(M)_{highest})$
Learning rate	0.1%
Epochs	100
Number of ANN neurons	2 hidden layers (32 and 16, respectively)
Activation functions used	ReLU in hidden layers SoftMax in final layer

The accuracy of the PCA–SRP-based ANN was compared against a model trained using ANN alone on different datasets.

5. Sc Range Identification

Accuracy testing was conducted for the specified dataset in order to identify the Sc range shown in Figure 9, to maximise the number of samples of the cleaned set S, and to maximise the number of samples of the removed set R in the processed $D_{PCA-SRP}$.
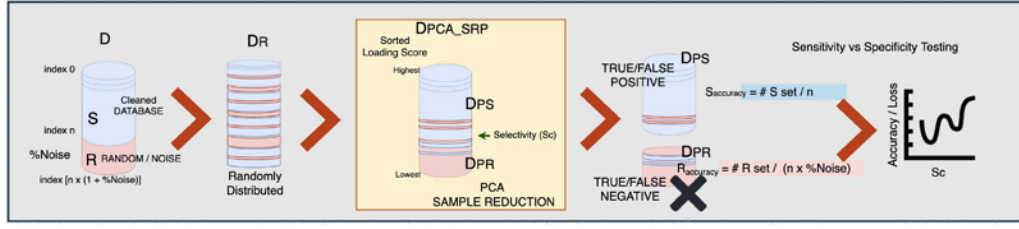
Figure 9. Detailed sensitivity vs. specificity testing procedure.

Assume that there are two sets of data D: cleaned data S and R.

$$D = S + R \qquad (5)$$

$$\%Noise = R/S \qquad (6)$$

$$R = \%Noise \times S \qquad (7)$$

Therefore,

$$D = S + \%Noise \times S \; D = S \times \qquad (8)$$

$$(1 + \%Noise) \qquad (9)$$

By indexing the dataset D for identifier, where n is the number of cleaned data S, we have the following:

$$D = \begin{cases} S & \text{if } 0 \leq \text{index} \leq n \\ R & \text{if } n < \text{index} \leq n \times (1 + \%Noise) \end{cases} \qquad (10)$$

D_R randomly distributes R in S to thoroughly mix the noise in the cleaned data by maintaining its index for identification. Then, D_R is processed in PCA–sample reduction
and sorted based on its loading score through the concept of covariance, eigenvalues and eigenvectors, and $D_{PCA-SRP}$.

Sc is the selectivity of the data based on the equation in Table 3, $D_{PCA-SRP}$, where the cut-off is biased to separate the clean-processed data D_{PS} and to remove processed data D_{PR}. The dataset was transformed from random to sorted PCASR-processed data.

Transformation:

$$D \rightarrow D_R \rightarrow D_{PCA-SRP} \qquad (11)$$

Thus,

$$D = S \cup R \qquad (12)$$

$$D_{PCA-SRP} = S_{PCA-SRP} \cup R_{PCA-SRP} \qquad (13)$$

By definition, $S_{accuracy}$ (true positive rate or sensitivity) is the number of correct samples in a clean-processed S in D_{PS} with a relative to number of n.

$$S_{accuracy} = S_{D_{PS}}/n \qquad (14)$$

However, $R_{accuracy}$ (True Negative Rate or Specificity) is the number of correct samples of R in D_{PR}.

$$R_{accuracy} = R_{DPR}/(n \times \%Noise)$$

(15)

Therefore, the most efficient *Sc* is where $S_{accuracy}$ and $R_{accuracy}$ meet.

5.1. Test Selectivity (Sc_{test})

To find Sc_{test}, the actual cleaned S set and random R set are needed to identify the accuracy by identifying the $S_{accuracy}$ and $R_{accuracy}$ lines and by applying Equation (16).

$$Sc_{test} = index[max(S_{accuracy} \times R_{accuracy})]$$

(16)

For Sc_{min}, $S_{accuracy}$ and $S_{accuracy}$ are maximised:

$$\frac{\partial(S_{accuracy})}{\partial(Sc)} = 0$$

(17)

$$\frac{\partial(R_{accuracy})}{\partial(Sc)} = 0$$

(18)

As shown in Equation (16),

$$F_{value} = S_{accuracy} \times R_{accuracy}$$

(19)

$$\frac{\partial F_{value}}{\partial(Sc)} = S_{accuracy} \times \frac{\partial(R_{accuracy})}{\partial(Sc)} + R_{accuracy} \times \frac{\partial(S_{accuracy})}{\partial(Sc)} = 0$$

(20)

$$\frac{\partial^2 F_{value}}{\partial(Sc)^2} < 0$$

(21)

Maximum F_{value} with respect to *Sc* is the Sc_{test} by satisfying the Equation (21).

Note: Be cautious when using Sc_{test} as *Sc*; it might remove all of the unwanted sample sets but may lose some information.

5.2. Minimum Selectivity (Sc_{min})

The minimum *Sc* is the rate of samples in PC_1, which shows that all included samples pass the minimum requirements of having the same attributes as the entire set.

$$Sc_{min} = PC_1 rate$$

(22)

Note: By using Sc_{min} as *Sc*, while it might not lose important information, the dataset *D* includes an abundance of random and noise samples.

The acceptable *Sc* is shown in Equation (23) in Figure 10.

$$0 \geq Sc_{min} \geq Sc \geq Sc_{test} \geq 100$$

(23)

Then,

$$Sc_{min} \geq Sc \geq Sc_{test} \geq 100$$

(24)

Therefore, the recommended Sc is between the range given in Equation (25). It varies depending on how much cleaning and removing are needed in the dataset D.
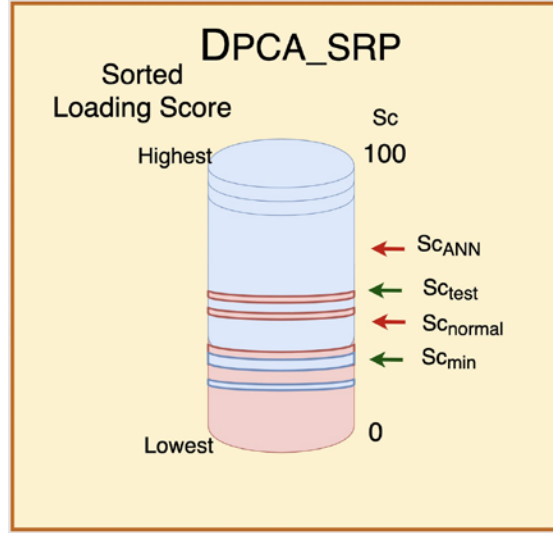
$$Sc_{min} \geq Sc \geq 100 \qquad (25)$$

Figure 10. Sc—accuracy testing representation.

5.3. ANN Selectivity (Sc_{ANN})

The ANN requires samples that allow for strong predictive power (see Section 7 for more details concerning the ANN requirements); we assumed the following:

$$Sc_{test} \geq Sc_{ANN} \geq 100 \qquad (26)$$

However, Sc_{test} is unknown due to unidentifiable R samples, and it is suggested that Sc_{ANN} is close to 100%.

6.   Multidimensional Datasets The multidimensional datasets for classification problems are tested under two layers of artificial neural network with 32 and 16 neurons, respectively, using ReLU and SoftMax activators, a testing size of 20%, a training size of 80%, 100 epochs, a learning rate of 10% [42], and Sc = 98% of the highest loading score in PCA analysis. Noisy data based on a standard distribution are added to the original dataset before training to simulate noisy data samples. Figure 11 shows how a given multidimensional dataset is pre-processed and used in the proposed PCA–SRP approach. Each multidimensional dataset or corpora has selected the variation in its dimensions, primarily sample sizes; the number of classifications shown in Table 4; the noise sample contents; and the purpose it serves in any field of biomedical applications in classification problems.

Table 4. Datasets used and their metadata.

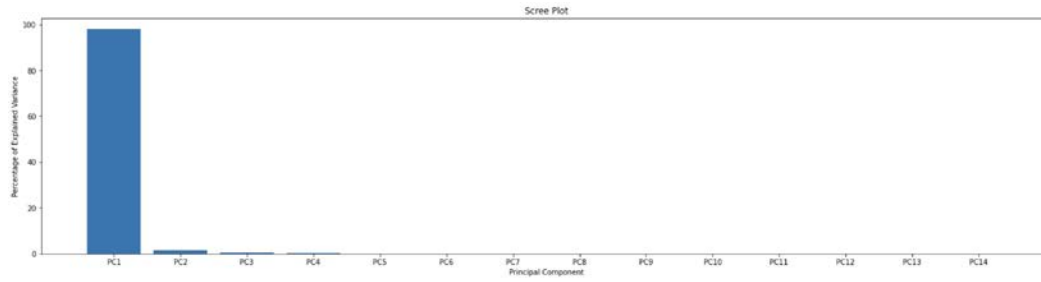| Datasets	Number of Dimensions	Sample Size	Target Classification
Heart Disease	14	301	2
Gender Voice Recognition	21	3167	2
Breast Cancer	31	568	2
Cancer Patients	24	1098	3
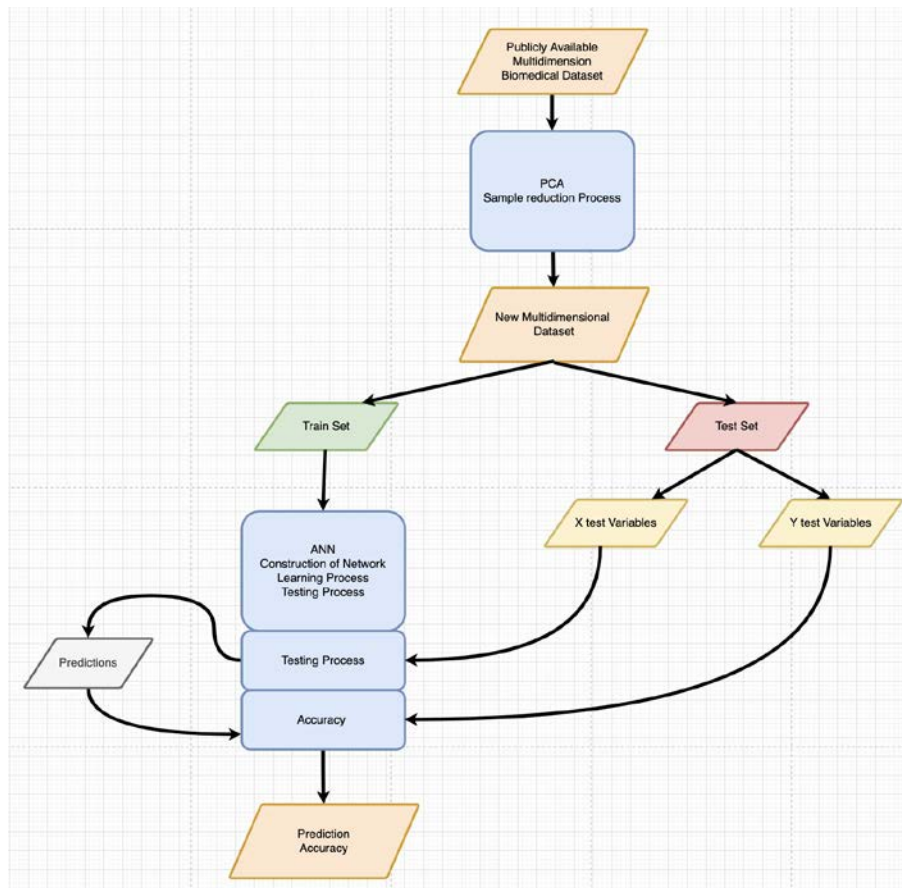
Figure 11. Pre-processing an application of multidimensional datasets in the proposed approach.

The different multidimensional datasets are shown in Table 4 and were acquired from different scientific and medical laboratories [43–47]. Specifically, to test our procedure, we vary the number of samples and dimensions. Moreover, researchers also are trying to show that the Python implementation works with a variety of dataset sizes.

6.1. Heart Disease Dataset

The experiment with the Cleveland patients dataset is concentrated on simply attempting to distinguish presence (values 1) from absence (values 0) to find other trends in heart data to predict certain cardiovascular events or to find clear indications of heart health [43].

Figures 12 and 13 show the result of the scree plot based on the values, indicating the PCA value distributions of the features in the dataset and the *Sc*-threshold cut-off limit, respectively. Based on the values, 78.48% of the top samples were selected in the training of the ANN model for a 98% selectivity.

**Figure 12.** Scree plot of the heart disease dataset.

the sorted samples based of loading scores

278	0.057987
226	0.057981
2	0.057976
45	0.057968
42	0.057961
...	...
147	0.003164
144	0.001178
138	0.000718
139	0.000605
327	0.000470

Length: 330, dtype: float64
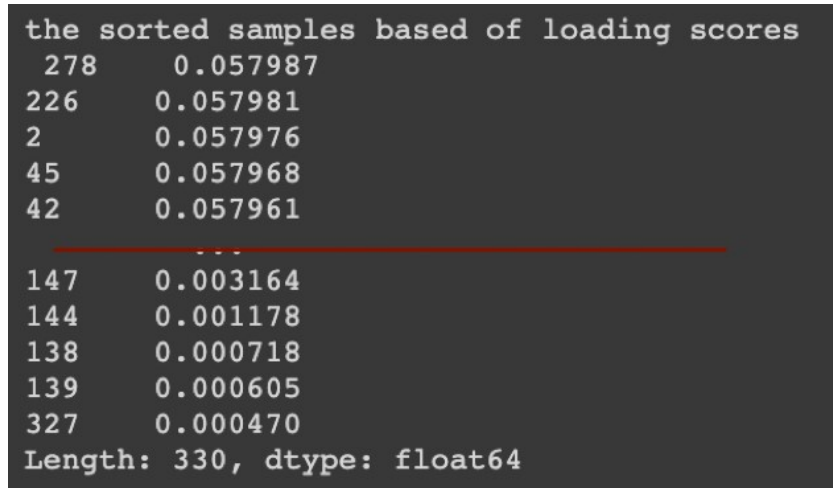
Figure 13. Threshold cutoff limit of the heart disease dataset.

6.2. Gender Voice Recognition Dataset

The gender voice recognition dataset is based upon acoustic properties of the voice and speech to identify a voice as male or female. The dataset consists of 3167 recorded voice samples collected from male and female speakers. The voice samples are pre-processed by acoustic analysis in R using the Seewave and TuneR packages, with an analysed frequency range of 0–280 hz (human vocal range) [44].

Figures 14 and 15 show the threshold cut-off limit and the result of the scree plot show the PCA distributions of the features in the dataset, respectively. In the training of the ANN model, the proposed method selected 23.82% of the top samples from the dataset.
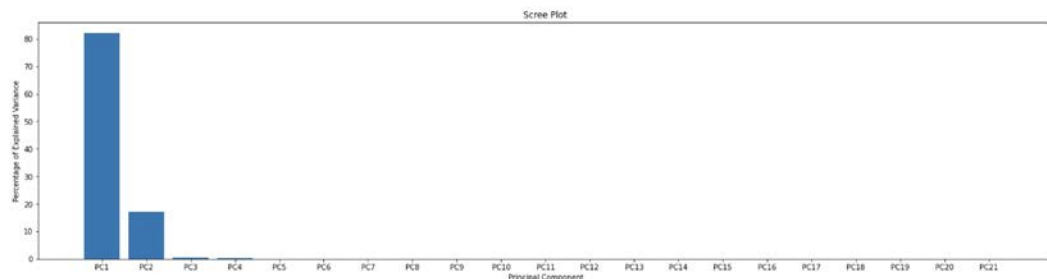
Figure 14. Scree plot of the gender voice recognition dataset.

```
the sorted samples based of loading scores
 2462     0.019580
3121      0.019578
2729      0.019575
2022      0.019574
2711      0.019571
            ...
1601      0.000244
1626      0.000243
3259      0.000181
1641      0.000168
3267      0.000068
Length: 3300, dtype: float64
```

Figure 15. Threshold cut off limit of the gender voice recognition dataset.

6.3. Breast Cancer Classification Dataset

The computed dataset consists of a series of digitised images of a mass breast fine needle aspirate (FNA). They describe characteristics of the cell nuclei present in the three-dimensional image as described in [45].

Figures 16 and 17 show the threshold cutoff limit and the results of the scree plot showing the PCA distributions of the features in the dataset, respectively. The selection of the top eigenvalues in the samples obtains 93.67% of the original dataset.
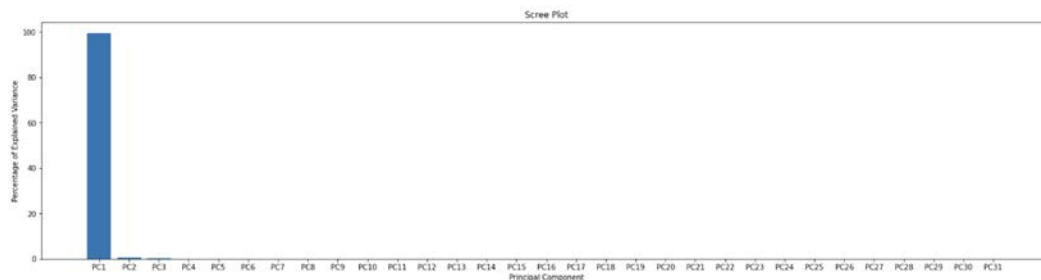
Figure 16. Scree plot of the breast cancer classification.

```
the sorted samples based of loading scores
 466      0.041997
545       0.041996
224       0.041996
438       0.041995
93        0.041995
            ...
580       0.001718
578       0.001383
594       0.000928
576       0.000828
595       0.000477
Length: 600, dtype: float64
```

Figure 17. Threshold cutoff limit of the breast cancer classification dataset.

6.4. Cancer Patients Dataset

The data comprise information about hundreds of cancer patients and their lifestyles. It consists of three classes (low, medium, and high severity), based on the cancer patients dataset [47].

Figures 18 and 19 provide the threshold cut-off limit and a graphical description of the PCA distributions amongst the features in the dataset, respectively. The selection of high eigenvalues features resulted in 45.41% of the samples in the original dataset used for the model's training.
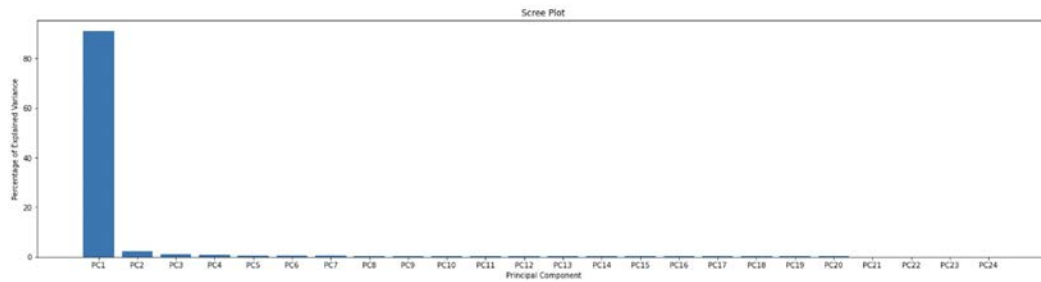
Figure 18. Scree plot of the cancer patients dataset.

```
the sorted samples based of loading scores
 900      0.031442
776      0.031442
332      0.031442
443      0.031442
110      0.031442
          ...
1082     0.006573
1061     0.006205
1043     0.004993
816      0.001869
98       0.000278
Length: 1099, dtype: float64
```
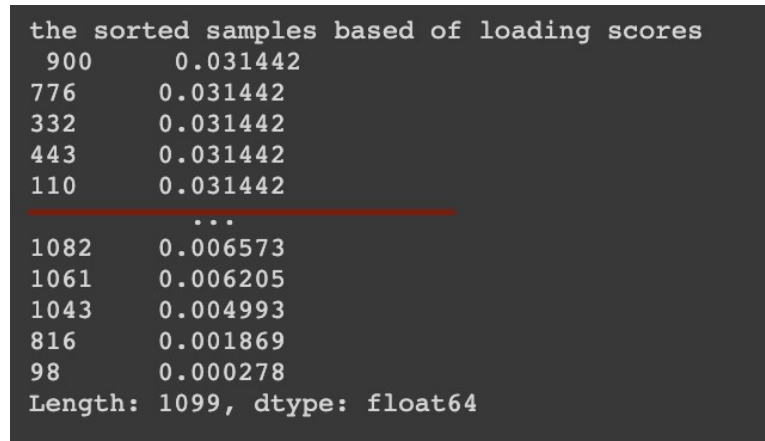
Figure 19. Threshold cut-off limit of the cancer patients dataset.

7. Discussion and Results

Upon acquiring the given datasets, they are contaminated with noise and random samples. By definition, noise is small, unwanted forms of energy or samples; on the other hand, random samples are the unconscious and unspecified values within the range of normal values.

The datasets with noise and random samples are processed using PCA–SRP with 98%selectivity (*Sc*), shown in Table 5, and then subjected to following tests:

- PCA–SRP + ANN comparison accuracy testing compares the validation model accuracy with and without PCA–SRP in an ANN.
- Sensitivity vs. specificity testing is a diagnostic test to find the approximate *Sc* range values (Sc_{test}).

- Receiver operating characteristic (ROC) curve testing compares the methodology PCA–SRP in different datasets in terms of organisation and classification of samples. Moreover, ROC curves also provide a practical evaluation of machine learning techniques.
- Accuracy vs. additional random samples testing is a diagnostic test responding to the sudden increase in noise and random sample spikes.

**Table 5.** Datasets sample status after PCA–sample reduction process.

| Dataset	Samples Used	% Pass rate
Heart Disease	259/330	78.48
Gender Voice Recognition	786/3300	23.82
Breast Cancer	562/600	93.67
Cancer Patients	499/1099	45.41

7.1. PCA–SRP + ANN Comparison Accuracy Testing

We compared the validity model accuracy using PCA–SRP in an ANN classification problem as suggested by the results presented in Figure 20 and determined its effect upon being subjected to noise and random samples. Tables 6 and 7 display the validation accuracy using ANN + PCA–SRP and ANN only, showing a significant increase as shown in Figure 21 for both noise and random samples.
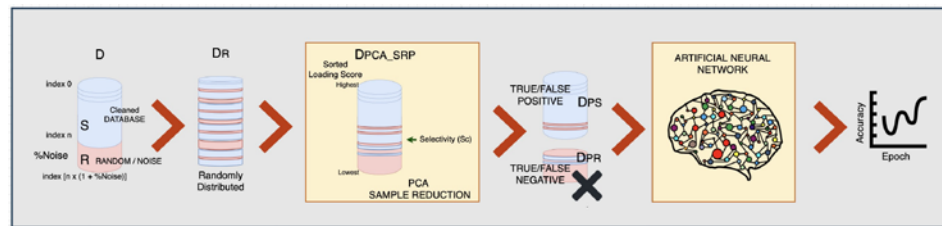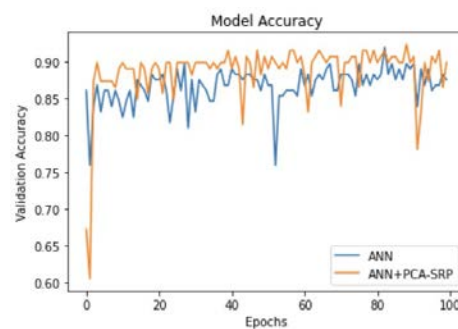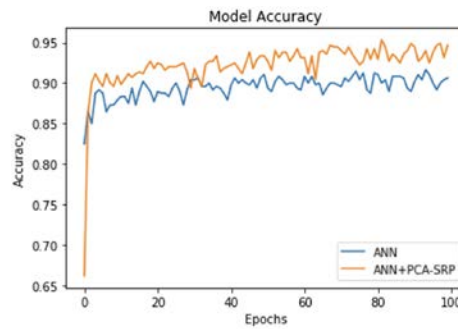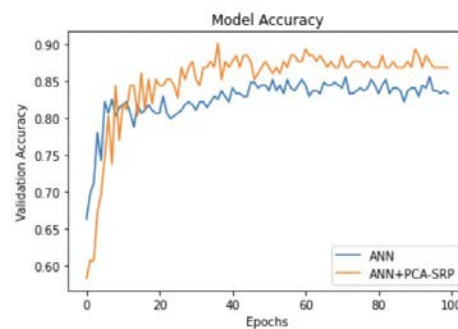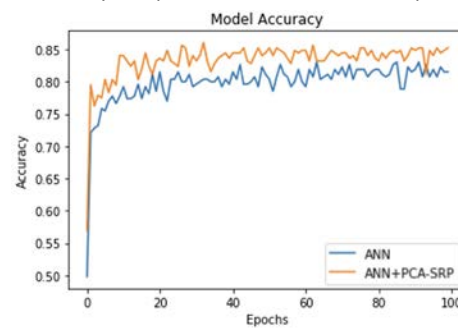

Figure 20. PCA–SRP to ANN process.
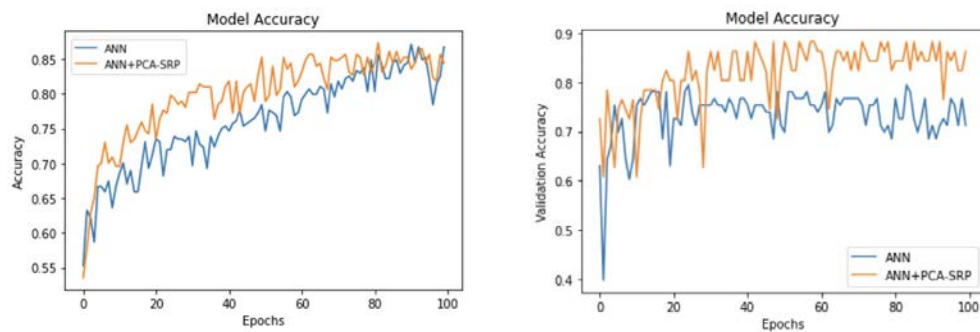
(**a**) Heart disease dataset (**b**) Heart disease dataset with noisy samples.
with random samples.
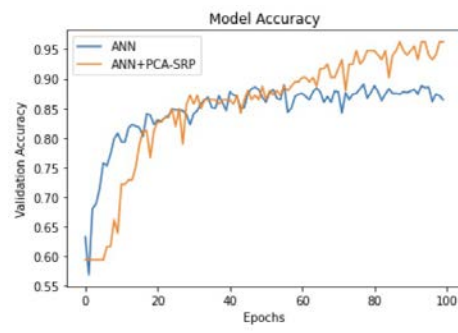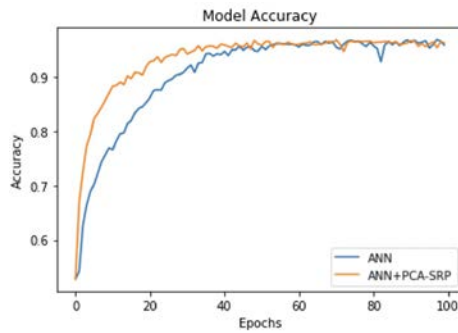
(**c**) Gender voice recognition dataset (**d**) Gender voice recognition dataset with noisy
samples.   with random samples.

(**e**) Breast cancer classification dataset   (**f**) Breast cancer classification dataset
with noisy samples.   in random samples.

(**g**) Cancer patients dataset   (**h**) Cancer patients dataset with noisy
samples.   with random samples.

Figure 21. Validation model accuracy of the given datasets.

The difference between dataset accuracy subjected to noise and random samples are the definition of the accuracy
 curve line between ANN + PCA–SRP to ANN only,

even though both of them have a significant accuracy increase. The accuracy of datasets under the influence of random
samples swings and deviates more and is less defined in comparison with accuracy with noise samples.

Table 6. Accuracy of datasets with noisy samples.

Dataset	ANN Only	ANN with SRP	% Increase
Heart Disease	76.19	79.89	3.7
Gender Voice Recognition	90.6	93.29	2.69
Breast Cancer	90.49	92.36	1.86
Cancer Patients	79.82	83.27	3.45

Table 7. Accuracy of datasets wit   random samples.

Dataset	ANN Only	ANN with SRP	% Increase
Heart Disease	73.69	81.76	8.07
Gender Voice Recognition	85.72	91.2	5.42
Breast Cancer	83.3	86.46	3.12
Cancer Patients	80.90	88.81	7.91

7.2. Sensitivity vs. Specificity Testing

Sensitivity measures how many true positives remain in S set, and it is described as a sudden dip as Sc increases. Specificity measures how many true negatives remain in the removed R set and increases as Sc increases. Figure 22 presents the diagnostic testing yielding a recommended Sc range in normal cleaning and ANN applications, as demonstrated in Equations (25) and (26), respectively.
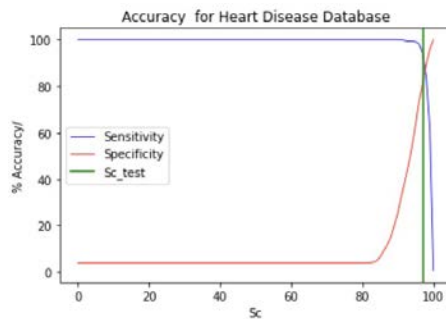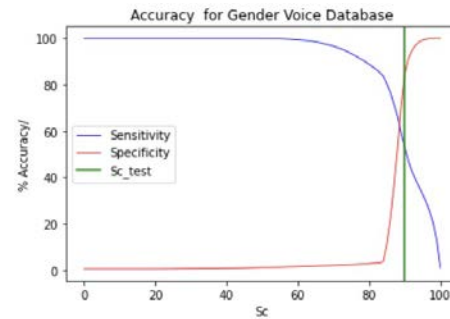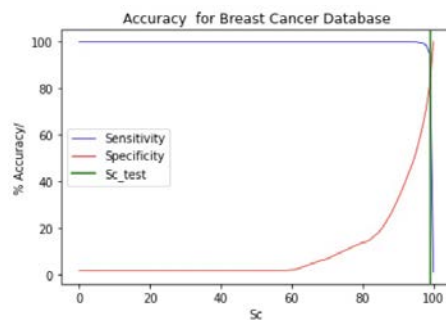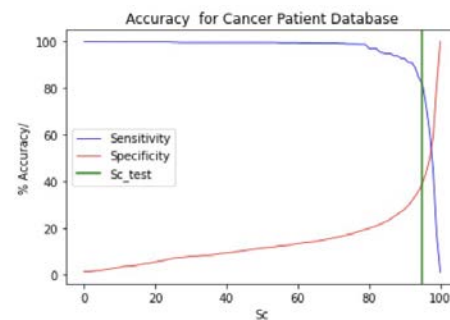

(**a**) Heart disease dataset.

(**b**) Gender voice recognition dataset.

(**c**) Breast cancer classification dataset.

(**d**) Cancer patients dataset.

Figure 22. Sensitivity vs. specificity diagram.

Table 8 shows the relation between Sc_{min} and Sc_{test}, but in practicality, Sc_{test} is hard to determine due to unknown R sets, and whether it consists of noise or random samples. In general, for the normal cleaning process, Sc is close to or above the Sc_{min} or PC_1 rate. For the given datasets, it is suggested that Sc_{normal} is as follows:

$$80\% \leq Sc_{normal} \leq 96\% \qquad (27)$$

Table 8. *Saccuracy ($Sc_{test-min}$)* result.

Dataset	Sc_{min}	Sensitivity	Specificity	Sc_{test}	Sensitivity	Specificity
Heart Disease	96	98.04	73.68	97	94.15	81.04
Gender Voice	80	88.75	2.87	90	53.08	84.103
Recognition Breast Cancer	96	99.47	57.64	99	94.51	81.84
Cancer Patients	88	93.83	25.86	95	81.8	39.1

However, ANN applications need highly predictive samples [12], so it is suggested that selectivity (Sc) should be almost near 100%. Table 8 shows Sc_{ANN} at the follow-ing range:

$$90\% \leq Sc_{ANN} \leq 99\% \qquad (28)$$

A high Sc value loses information and true positives in the S set but increases the specificity of the dataset; however, a low value of Sc adds noise and random samples that yield fewer true negatives and increases sensitivity. Careful adjustment of Sc vouches for a good result as a cleaning agent in the system.
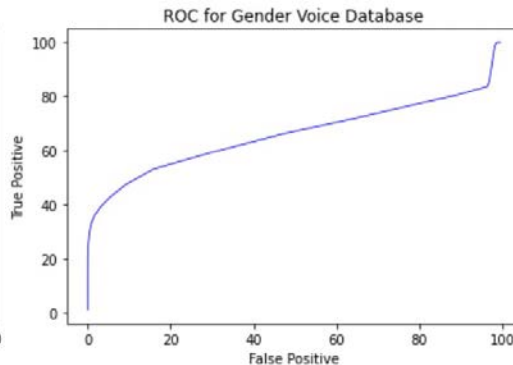
7.3. Receiver Operating Characteristic (ROC) Curve Testing

As observed in Figure 23, receiver operating characteristic (ROC) datasets have shown an organisation of the samples, even the random and noise samples were strongly mixed into the cleaned set. The larger the area under the curve (AUC), the better the classifier methodology for the true positive rate (sensitivity) vs. true negative rate (1 − specificity) diagram, as seen in Figure 24. Ideally, the objective is the perfect classifier; nevertheless, a result above the random classifier line would allow us to conclude that the methodology is acceptable.
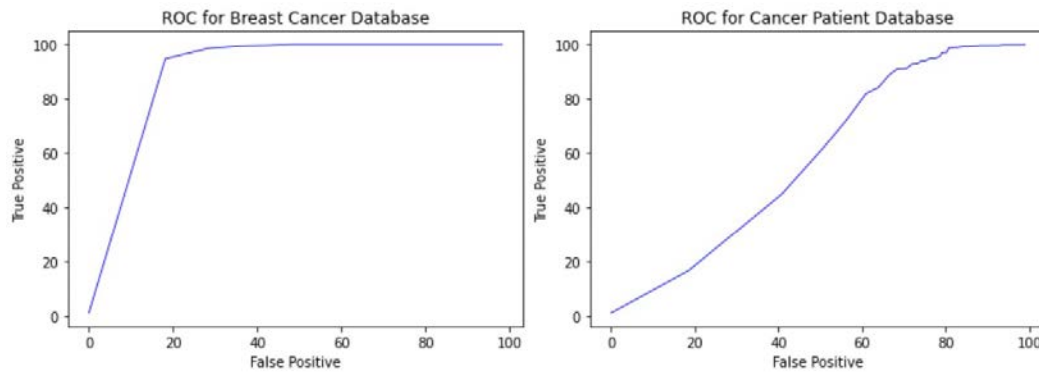
Given all of the datasets, the cancer patients dataset is acceptable but by not as much as the other datasets; it has a smaller AUC.
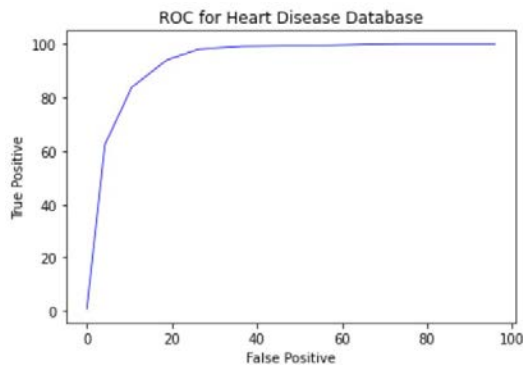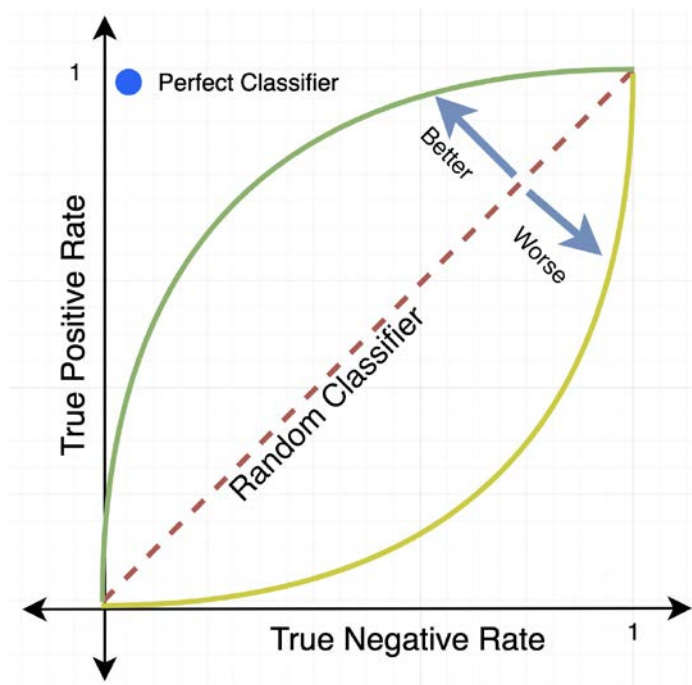

(**a**) Heart disease dataset.   (**b**) Gender voice recognition dataset.

(**c**) Breast cancer classification dataset.   (**d**) Cancer patients dataset.

Figure 23. Receiver operating characteristic (ROC) curve.

Figure 24. ROC curve interpretation.

7.4. Accuracy vs. Additional Random Samples Testing

Since the cancer patients dataset shows the least-effective classifier, as shown in Figure 23, it was tested for its response to the injection of a sudden spike in additional random samples up to 100%. Figure 25 shows that, upon increasing the random samples, the result of PCA–SRP decreases gradually.

Figures 26 and 27 present the validation accuracy of both ANN + PCA(SRP) and ANN only to additional noises up to 100% with selectivities (Sc) of 88 and 98%, respectively. It has also shown a reasonable increase in accuracy using 98% selectivity (*Sc*) instead of 88%.
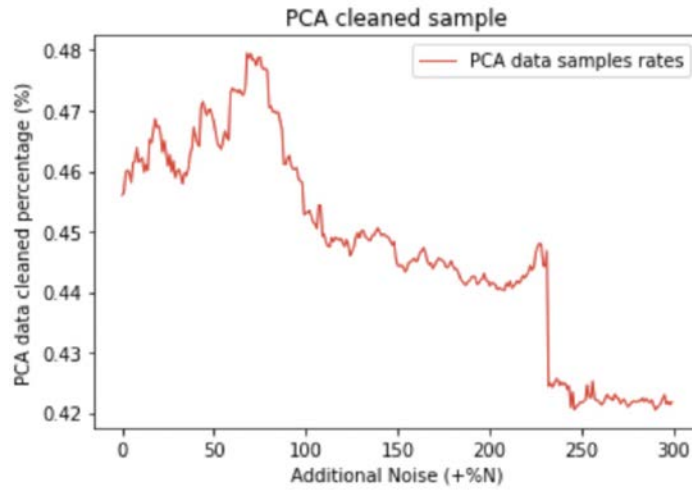
Figure 25. PCA cleaned data percentage (%).

The high-valued data have been preserved and maintain their accuracy until a specific additional noise point. Nevertheless, the predictive power remains intact until that point, even though some data were lost in the process.
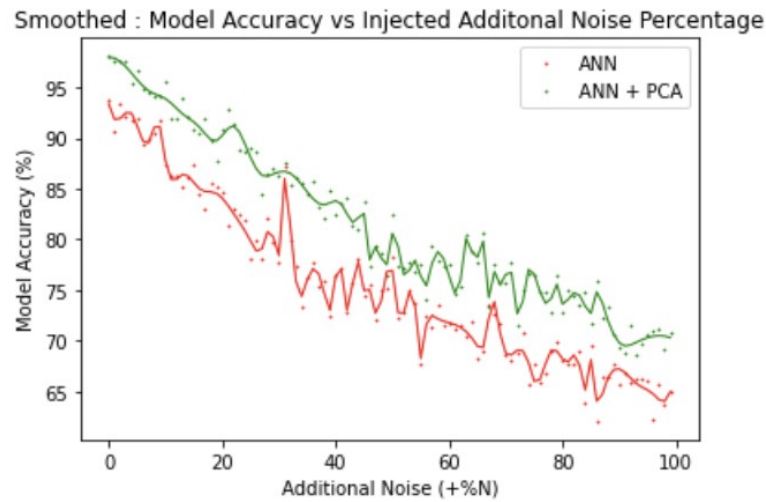
Figure 26. +%Noise response for cancer patients dataset (Sc = 88%).

The *ANN* $^+$ *PCA$_{SRP}$* maintains the highest performance of cancer patient classifications with high *Sc* values. The methodology allows for a significant advantage by gradually lessening the decrease in classification problem accuracy over the sudden increase in noise in the system.
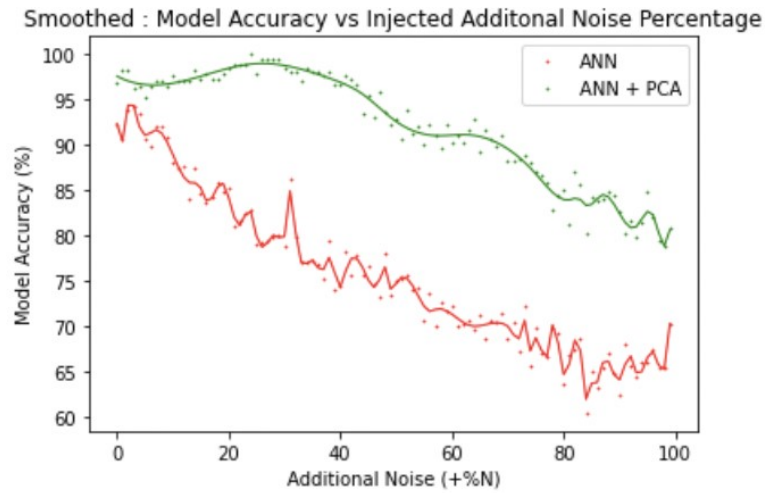
Figure 27. +%Noise response for the cancer patients dataset (Sc = 98%).

The ANN classification problem requires strong training sets, which requires a lot of high *Sc* while disregarding the low ones; based on the observation in both Figures 26 and 27, Sc_{ANN} is better, as described in Equation (26).

8. Conclusions and Future Research The material presented in this paper shows a significant improvement in the accuracy of an ANN in classification problems with the aid of principal component analysis–sample reduction process (PCA–SRP). The ANN cast off 10% of the learning rate, two layers with 32 and 16 hidden neurons, ReLU activators in hidden layers, and SoftMax in output activators with 100 epochs or iterations, as shown in Table 3 based on the PCA–SRP and ANN The Python implementation program was implemented on the multidimensional datasets gathered, namely the heart disease, gender voice recognition, breast cancer classification, and cancer patients datasets provided in Table 4. These datasets were then used in the PCA–SRP + ANN accuracy comparison testing, sensitivity vs. specificity testing, receiver operating characteristic (ROC) curve testing, and accuracy vs. additional random samples testing; the results show significant improvements.

PCA–SRP removed dirty and imprecise datasets, based on the results shown in Table 5, which allowed us to reduce the number of samples in the process and allowed for a significant increase in accuracy, as shown in Tables 6 and 7. Furthermore, we also determined the recommended *Sc* range values for normal cleaning and the ANN classification problem.

Future research will further investigate the performance in massive biomedical datasets and determine how to load them into PCA–SRP cleansing agents; one of the suggestions is loading through batch processing. Another suggestion is to utilise knowledge-based techniques of PCA–SRP in different neural network architectures such as CNNs, RNNs, LSTMs, and GNNs. Furthermore, an investigation into various field applications to explore the incorporation of the investigated cleaning techniques, such as real-time biomedical automation, image-based medical diagnosis classification, and human thoughts processes [48–51] could be a desirable research avenue.

The proposed methodology can be applied to a wearable EEG or similar device in order to extract chaotic data from the brain's unique biometric featured samples for use in cryptography or especially for steganography. Lastly, this formal basis is necessary to design and construct high-quality and helpful software tools to support the data cleansing process of PCA–SRP and its application for artificial neural networks (ANNs).

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
CNN	Convolutional Neural Network
GNN	Graph Neural Network
LSTM	Long Short-Term Memory
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SCARE	Scalable Automatic Repairing
SRP	Sample Reduction Process

References

1. Engin, M.; Demirel, A.; Engin, E.Z.; Fedakar, M. Recent developments and trends in biomedical sensors. *Measurement* **2005**, *37*, 173–188. [CrossRef]
2. Hannan, M.A.; Abbas, S.M.; Samad, S.A.; Hussain, A. Modulation Techniques for Biomedical Implanted Devices and Their Challenges. *Sensors* **2012**, *12*, 297. [CrossRef] [PubMed]
3. Gholami, M.; Cai, N.; Brennan, R. An artificial neural network approach to the problem of wireless sensors network localization. *Robot. -Comput.-Integr. Manuf.* **2013**, *29*, 96–109. [CrossRef]
4. Cheng, L.; Guan, Y.; Zhu, K.; Li, Y. Recognition of human activities using machine learning methods with wearable sensors. In Proceedings of the 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 9–11 January 2017; pp. 1–7.
5. Barnett, M.P. The Information Explosion. *Nature* **1964**, *203*, 585–585. [CrossRef]
6. Schwiebert, L.; Gupta, S.K.S.; Weinmann, J. Research challenges in wireless networks of biomedical sensors. In Proceedings of the 7th Annual International Conference on Mobile Computing and Networking, Rome, Italy, 16 July 2001; pp. 151–165.
7. Fung, B.C.; Wang, K.; Yu, P.S. Top-down specialization for information and privacy preservation. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; pp. 205–216.
8. Zhang, L.; Wu, X. Recent Progress in Challenges of Wireless Biomedical Sensor Network. In Proceedings of the 2009 3rd International Conference on Bioinformatics and Biomedical Engineering, Beijing, China, 11–13 June 2009; pp. 1–4.
9. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
10. Geiger, B.C.; Kubin, G. Relative information loss in the PCA. In Proceedings of the 2012 IEEE Information Theory Workshop, Lausanne, Switzerland, 3–7 September 2012; pp. 562–566.
11. Rahm, E.; Do, H.H. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* **2000**, *23*, 3–13.
12. Ridzuan, F.; Wan Zainon, W.M.N. A Review on Data Cleansing Methods for Big Data. *Procedia Comput. Sci.* **2019**, *161*, 731–738. [CrossRef]
13. Wang, H.; Li, M.; Bu, Y.; Li, J.; Gao, H.; Zhang, J. Cleanix: A big data cleaning parfait. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai, China, 3–7 November 2014; pp. 2024–2026.
14. Yakout, M.; Berti-Équille, L.; Elmagarmid, A.K. Don't be scared: Use scalable automatic repairing with maximal likelihood and bounded changes. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 22–27 June 2013; pp. 553–564.
15. Chu, X.; Morcos, J.; Ilyas, I.F.; Ouzzani, M.; Papotti, P.; Tang, N.; Ye, Y. KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, VIC, Australia, 31 May–4 June 2015; pp. 1247–1261.
16. Khayyat, Z.; Ilyas, I.F.; Jindal, A.; Madden, S.; Ouzzani, M.; Papotti, P.; Quiané-Ruiz, J.-A.; Tang, N.; Yin, S. Bigdansing: A system for big data cleansing. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, VIC, Australia, 31 May–4 June 2015; pp. 1215–1230.

17. Chu, X.; Ilyas, I.F.; Krishnan, S.; Wang, J. Data Cleaning: Overview and Emerging Challenges. In Proceedings of the Proceedings of the 2016 International Conference on Management of Data, San Francisco, CA, USA, 26 June–1 July 2016; pp. 2201–2206.

18. Badri, L. Development of neural networks for noise reduction. *Int. Arab J. Inf. Technol.* **2010**, *7*, 289–294.

19. Kalantari, N.K.; Bako, S.; Sen, P. A machine learning approach for filtering Monte Carlo noise. *ACM Trans. Graph.* **2015**, *34*, 122-1. [CrossRef]

20. Kim, K.H.; Kim, S.J. Neural spike sorting under nearly 0-dB signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 1406–1411.

21. Maletic, J.I.; Marcus, A. *Data Cleansing: Beyond Integrity Analysis*; Pennsylvania State University: University Park, PA, USA, 2000; pp. 200–209.

22. Jilani, T.A.; Yasin, H.; Yasin, M.M. PCA-ANN for classification of Hepatitis-C patients. *Int. J. Comput. Appl.* **2011**, *14*, 1–6. [CrossRef]

23. Deshpande, N.T.; Ravishankar, S. Face Detection and Recognition using Viola-Jones algorithm and Fusion of PCA and ANN. *Adv. Comput. Sci. Technol.* **2017**, *10*, 1173–1189.

24. Uğˇuz, H. A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases. *J. Med Syst.* **2012**, *36*, 61–72. [CrossRef]

25. Narendra, N.; Alku, P. Dysarthric speech classification from coded telephone speech using glottal features. *Speech Commun.* **2019**, *110*, 47–55. [CrossRef]

26. Moreira, L.P.; Silveira, L.; Pacheco, M.T.T.; da Silva, A.G.; Rocco, D.D.F.M. Detecting urine metabolites related to training performance in swimming athletes by means of Raman spectroscopy and principal component analysis. *J. Photochem. Photobiol. B Biol.* **2018**, *185*, 223–234. [CrossRef] [PubMed]

27. Benhar, H.; Idri, A.; Fernández-Alemán, J. Data preprocessing for heart disease classification: A systematic literature review. *Comput. Methods Programs Biomed.* **2020**, *195*, 105635. [CrossRef]

28. Kara, S.; Güven, A.; Içer, S. Classification of macular and optic nerve disease by principal component analysis.˙ *Comput. Biol. Med.* **2007**, *37*, 836–841. [CrossRef]

29. Krisciukaitis, A.; Tamosiunas, M.; Jakuska, P.; Veteikis, R.; Lekas, R.; Saferis, V.; Benetis, R. Evaluation of ischemic injury of the cardiac tissue by using the principal component analysis of an epicardial electrogram. *Comput. Methods Programs Biomed* **2006**, *82*, 121–129. [CrossRef]

30. Kumar, V.; Sachdeva, J.; Gupta, I.; Khandelwal, N.; Ahuja, C.K. Classification of brain tumors using PCA-ANN. In Proceedings of the 2011 World Congress on Information and Communication Technologies, Mumbai, India, 11–14 December 2011; pp. 1079–1083. 31. Walse, K.H.; Dharaskar, R.V.; Thakare, V.M. PCA Based Optimal ANN Classifiers for Human Activity Recognition Using Mobile Sensors Data. In Proceedings of the First International Conference on Information and Communication Technology for Intelligent Systems, Ahmedabad, India, 28–29 November 2015; Volume 1, pp. 429–436.

32. Landi, A.; Piaggi, P.; Pioggia, G. Backpropagation-based non linear pca for biomedical applications. In Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications, Pisa, Italy, 30 November–2 December 2009; pp. 635–640.

33. Mostaar, A.; Sattari, M.; Hosseini, S.; Deevband, M. Use of Artificial Neural Networks and PCA to Predict Results of Infertility Treatment in the ICSI Method. *J. Biomed. Phys. Eng.* **2019**, *9*, 679. [CrossRef]

34. Mehta, D.; Verma, C. Prediction of cancer diagnosis patients from fine-needle aspirates using machine learning. In Proceedings of the International Conference on Intelligent Computing and Smart Communication, Tehri, India, 20–21 April 2019; pp. 337–348.

35. Chen, Y.; Hao, Y. Integrating principle component analysis and weighted support vector machine for stock trading signals prediction. *Neurocomputing* **2018**, *321*, 381–402. [CrossRef]

36. Zhong, X.; Enke, D. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financ. Innov.* **2019**, *5*, 1–20. [CrossRef]

37. Adisa, J.A.; Ojo, S.O.; Owolawi, P.A.; Pretorius, A.B. Financial Distress Prediction: Principle Component Analysis and Artificial Neural Networks. In Proceedings of the 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Vanderbijlpark, South Africa, 21–22 November 2019; pp. 1–6.

38. Tran, N.M.; Burdejová, P.; Ospienko, M.; Härdle, W.K. Principal component analysis in an asymmetric norm. *J. Multivar. Anal.* **2019**, *171*, 1–21. [CrossRef]

39. Richards, J.E. Recovering dipole sources from scalp-recorded event-related-potentials using component analysis: Principal component analysis and independent component analysis. *Int. J. Psychophysiol.* **2004**, *54*, 201–220. [CrossRef]

40. Chen, Z.; Oni, S.; Hoban, S.; Jademi, O. A Comparative Study of Data Cleaning Tools. *Int. J. Data Warehous. Min.* **2019**, *15*, 48–65. [CrossRef]

41. Dubey, A. The Mathematics Behind Principal Component Analysis. Available online: https://tinyurl.com/PCA-Maths (accessed on 12 April 2020).

42. Agarap, A.F. Deep learning using rectified linear units (ReLU). *arXiv* **2018**, arXiv:1803.08375.

43. Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Detrano, R. Heart Disease Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/Heart+Disease (accessed on 1 April 2019).

44. Becker, K. Gender Recognition by Voice. Available online: https://www.kaggle.com/primaryobjects/voicegender (accessed on 1 February 2020).

45. Wolberg, W.H.a.S.; William, N.; Mangasarian, O.L. Breast Cancer Wisconsin (Diagnostic) Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/BreastCancerWisconsin(Diagnostic) (accessed on 12 November 2020).

46. Barreto, G.d.A.; Neto, A. R.; Filho, H. A. Vertebral Column Data Set. Available online: http://archive.ics.uci.edu/ml/datasets/VertebralColumn (accessed on 1 August 2019).

47. Da Marla, R. Cancer Patients Data Set. Available online: https://data.world/cancerdatahp/lung-cancer-data/workspace (accessed on 1 March 2020).

48. Vyas, S.; Das, S.; Ting, Y.-P. Predictive modeling and response analysis of spent catalyst bioleaching using artificial neural network. *Bioresour. Technol. Rep.* **2020**, *9*, 100389. [CrossRef]

49. Yalung, C.N.; Adolfo, C.M.S. Analysis of obstacle detection using ultrasonic sensor. *Int. Res. J. Eng. Technol.* **2017**, *4*, 1015–1019.

50. Tirupattur, P.; Rawat, Y.S.; Spampinato, C.; Shah, M. Thoughtviz: Visualizing human thoughts using generative adversarial network. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, South Korea, 22–26 October 2018; pp. 950–958.

51. Yalung, C.; Majeed, S.A.; Adolfo, C.M.; Karam, J.; Mirtskhulava, L. Gyroscope explorer terrain angles classification. In Proceedings of the 2016 IEEE East-West Design & Test Symposium (EWDTS), Yerevan, Armenia, 14–17 October 2016; pp. 1–5.