Received: 8 July 2022 | Revised: 20 September 2022 | Accepted: 22 September 2022 | Published online: 30 September 2022

RESEARCH ARTICLE

Artificial Intelligence and Applications
2023, Vol. 1(1) 35–42
DOI: 10.47852/bonviewAIA2202303

BON VIEW PUBLISHING

Toward a Self-Supervised Architecture for Semen Quality Prediction Using Environmental and Lifestyle Factors

Ejay Nsugbe[1],*

[1]*Nsugbe Research Labs, UK*

Abstract: Male fertility has been seen to be declining, prompting for more effective and accessible means of its assessment. Artificial intelligence methods have been effective toward predicting semen quality through a questionnaire-based information source comprising a selection of factors from the medical literature which have been seen to influence semen quality. Prior work has seen the application of supervised learning toward the prediction of semen quality, but since supervised learning hinges on the provision of data class labels it can be said to depend on an external intelligence intervention, which can translate toward further costs and resources in practical settings. In contrast, unsupervised learning methods partition data into clusters and groups based on an objective function and do not rely on the provision of class labels and can allow for a fully automated flow of a prediction platform. In this paper, we apply three unsupervised learning models with different model architectures, namely Gaussian mixture model (GMM), K-means, and spectral clustering (SC), alongside low dimensional embedding methods which include sparse autoencoder (SAE), principal component analysis (PCA), and robust PCA. The best results were obtained with a combination of the SAE and the SC algorithm, which was likely due to its nonspecific and arbitrary cluster shape assumption. Further work would now involve the exploration of similar unsupervised learning algorithms with a similar framework to the SC to investigate the extent to which various clusters can be learned with maximal accuracy.

Keywords: unsupervised learning, machine learning, decision support, public health, dimensionality reduction, bioinformatics

1. Introduction

As male fertility has been reported to be on the decline, due to a multitude of reasons, the ability to assess the quality of a male semen sample is emphasized as a means of identifying potential fertility disorders (Grant et al., 2006; Inhorn, 2003). The typical assessment for semen quality involves the traditional laboratory-based test where fertility metrics are tested according to the guidelines provided by the World Health Organization (WHO). Despite the effectiveness of this method, it is known that avenues for uncertainties exist that creep into the reporting of the results due to subjectivity involved in the interpretation of measurement and general errors (Tomlinson, 2016). Furthermore, it has been acknowledged that the cost of the acquisition of the relevant instrumentation for the purpose of semen analysis, coupled with the price for a standard semen test, has been viewed to be typically unaffordable in developing economies due to a combination of constrained healthcare funding alongside low incomes, which in turn renders this means of semen analysis to be inaccessible to a bulk of the inhabitants within that region (Gerrits & Shaw, 2010).

Pattern recognition and predictive models have shown in prior case studies to be able to predict aspects of overall semen quality from a qualitative data source comprising of selected questions which have been seen to influence overall semen quality, as per the medical literature (Bidgoli et al., 2015; Levine et al., 2017). A recent work in this area by the authors applied a combination of low dimensional embedding methods alongside supervised learning models for prediction of semen quality (Nsugbe et al., Unpublished results). The inclusion of low dimensional embedding methods served as a preprocessing method that can help correct for anomalous entries into the questionnaire during the data gathering process, which could stem from a variety of reasons, most notably due to illiteracy in the case of developing economies, with the principal component analysis (PCA) and the robust PCA (RPCA) proving to be the more effective low dimensional embedding methods (Brunton, 2020; Provost, 2014; Verner, 2005).

As the supervised learning method was used for the classification of the semen quality into various classes, it can be noted that the process relies on labeling of the samples, which requires an external human intervention within the process (IBM Cloud Education, 2021; Kotsiantis, 2007). From a clinical perspective, a nonautomated process within the model prediction loop would likely demand additional resources, which may accumulate overtime and make the process unappealing, particularly to low-income countries that can be assumed to operate within a resource-constrained clinical setting (Butler, 2015).

This stride toward the evolution of the intelligence level of the prediction model carries financial and resource-based appeal due to the framework of operation; thus, the emphasis of this paper is around the design of prediction models which can work in a fully automated fashion with minimal human intervention and thus

*Corresponding author: Ejay Nsugbe, Nsugbe Research Labs, UK. Email: ennsugbe@yahoo.com

represent a different variant of the artificial intelligence (AI) system from prior work (Bishop, 2002). To achieve this, the unsupervised learning framework was adopted as a means of enhancing the system autonomy and therein contributes toward enhancing the overall intelligence of the prediction model (Adorf et al., 2019; Alloghani et al., 2020). Thus, in this paper, we address the following:

- Application of a combination of low dimensional embedding methods alongside unsupervised learning methods for semen quality prediction
- A comparison between the various kinds of unsupervised learning approaches to pinpoint the unsupervised learning method best suited to data acquired from a questionnaire format.

2. Materials

2.1. Sperm production

The development, origin, and creation of sperm cells within the male testes are termed as spermatogenesis, where sperm production takes place within a set of miniature seminiferous tubes within the testes (Neto et al., 2016; Sharpe, 2010). Embedded inside the walls of the tubes are Sertoli cells, which ensure the continuous nourishment of the sperm cells through constant provision of blood and nutrients (Neto et al., 2016; Sharpe, 2010). Once the sperm cells are mature they are transported from the long seminiferous tubules and hosted within the epididymis, and are at the ready for when sufficient stimulation is produced to cause for them to leave the body via an ejaculation process (Neto et al., 2016; Sharpe, 2010). An annotated image showing an overview of the testis and the various stages associated with sperm production can be seen in Figure 1.

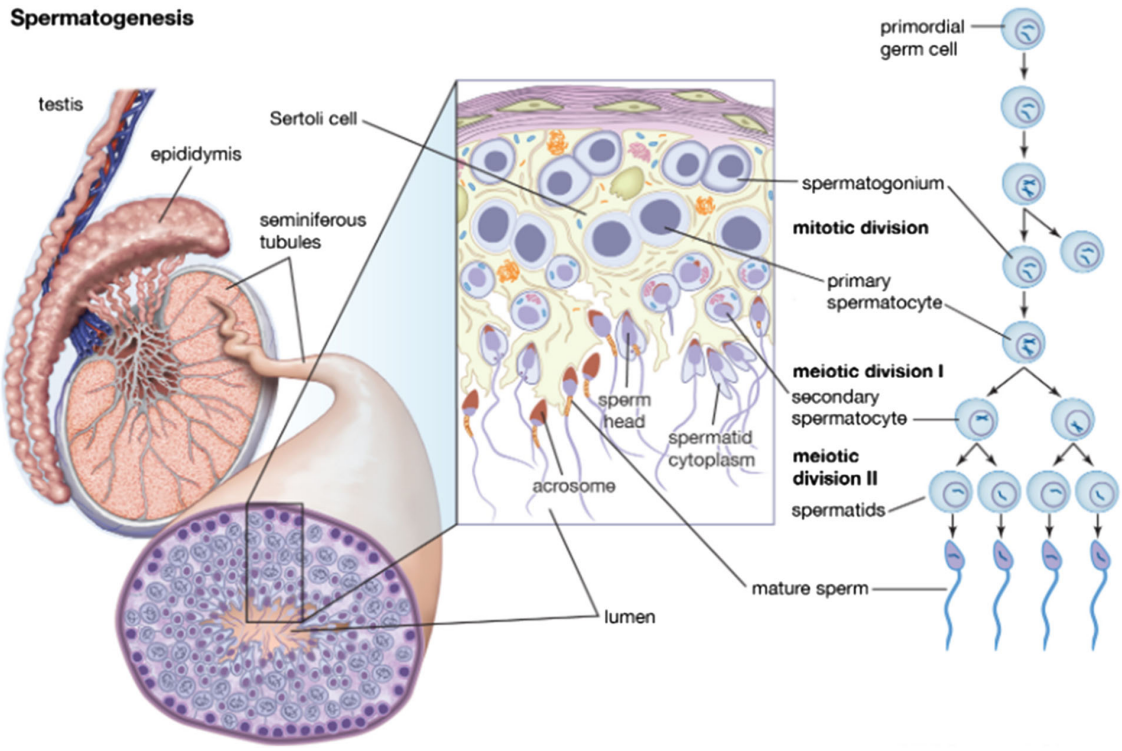
The data used for this paper are a qualitative feature vector from the UCI data repository containing data from 100 volunteers which were analyzed according to the baseline reference set by the WHO 2010 standards (Gil et al., 2012; *UCI Machine Learning Repository: Fertility Data Set*, n.d.). A combination of these select variables within the feature vector have been seen from a medical perspective to allow for a means toward predicting sperm-based quality factors such as the concentration from which a further inference can be carried out to infer sperm count, with the reason for this being that sperm concentration reflects the amount of sperm quantity present within a reference volume (Andrade-Rocha, 2013; Carlsen et al., 2003; Gunes et al., 2016; Hadi et al., 1987; Jørgensen et al., 2001; Jung et al., 2001; Levine, 1999; Sergerie et al., 2007; Sharma et al., 2015; Vignera et al., 2012; Your Fertility, n.d.).

Of the 100 samples downloaded from the database, 88 were labeled as normal with the other 12 labeled as altered, therein making it apparent that a class imbalance exists—this was remedied using the SMOTE artificial sample generator to contribute toward sample generation and class balancing (Blagus & Lusa, 2013).

2.2. Qualitative feature vector

The qualitative feature vector comprised the following 9 features, which have been seen in the medical literature to give an indication of semen quality: (1) Season—research has shown that semen quality varies by up to 30% between seasons, with the lowest quality being prevalent in the summer months, whereas the highest quality has been seen during the winter period (Nsugbe et al., Unpublished results). (2) Age—it is well accepted that aging primarily causes irreversible physiological changes in human beings, these changes manifest themselves on a cellular level where they are believed to have effects on the reproductive system and semen quality as a

Figure 1
An overview of the testis and the process of sperm production (Encyclopedia Britannica, n.d.)



whole (Nsugbe et al., Unpublished results). Several reports from the medical literature have also noted the occurrence of congenital birth issues in newborns, which range from neurological ailments to heart defects, and whom have fathers whose age is above the 40-year mark (Nsugbe et al., Unpublished results). (3) Childhood diseases—there has been shown to be links between different kinds of childhood disease and male reproductive issues which ultimately influence fertility (Nsugbe et al., Unpublished results). (4) Prior accident (non-testicular)—although strong correlations are yet to be formed between prior accident and semen quality, this factor was included in the final feature vector (Nsugbe et al., Unpublished results). (5) Surgical intervention (non-testicular)—as with the earlier feature on prior accidents, there continues to be medical research to find a concrete link between surgical interventions and semen quality, although this factor was also included in the final feature vector (Nsugbe et al., Unpublished results). (6) High fevers in the last 12 months—illnesses have shown to distort the spermatogenesis phase when semen is created, and as a result can influence the overall semen quality (Nsugbe et al., Unpublished results). (7) Alcohol consumption—a handful of studies have found correlations between the volume of alcohol consumption and respective semen quality (Nsugbe et al., Unpublished results), although this continues to be a growing area of research in the medical literature. (8) Smoking habits—the presence of polycyclic aromatic hydrocarbons has been seen to influence a proapoptotic protein which has a number of negative effects on male fertility (Nsugbe et al., Unpublished results). (9) Hours spent sitting per day—sitting for long periods has been seen to generate heat in the testicular region and in turn gives rise to oxidative stress which causes damage to molecules (Nsugbe et al., Unpublished results).

Table 1 shows the various features, values, and normalized ranges.

3. Proposed Methodology

3.1. Data preprocessing methods

During the preprocessing stage, the values inputted in the feature vector were cleansed to minimize the influence of errors and uncertainties during the model design stage. The preprocessing stage involved the application of three dimensionality reduction methods whose respective performance was compared and contrasted. In a previous study, the results showed that linear variants of dimensionality reduction methods supersede their nonlinear counterparts, with the PCA and RPCA in particular providing the model performance, relatively speaking (Nsugbe et al., Unpublished results; Raj, 2019; Zhou et al., 2010). In this paper, we employed both the PCA and RPCA in addition to the sparse autoencoder (SAE) for preprocessing and dimensionality reduction. A description of each method can be seen as follows:

Table 1
Table showing the various features, their description, and value range

Feature description	Values	Normalized scale
Season	1. Winter, 2. Spring, 3. Summer, 4. Autumn	−1, 0.33, 0.33, 1
Age range	18–36	0–1
Childhood disease	Yes or no	0, 1
Prior accident	Yes or no	0, 1
Surgical intervention	Yes or no	0, 1
High fevers in the last year	Less than 3 months, over 3 months, none	−1, 0, 1
Alcohol consumption	Several times a day, every day, several times a week, once a day, never/hardly ever	0–1
Smoking habit	Never, occasional and daily	−1, 0, 1
Hours spent sitting per day	0–16	0–1

PCA: a data compression algorithm that projects the data from a high dimensional representation to a lower dimensional structure by a means of linear combinations of an input feature vector (Jolliffe & Cadima, 2016; Raj, 2019). It is viewed as an unsupervised algorithm since associated data labels are not required as the algorithm effectively hinges its solution convergence on its ability to compute the solution to an eigenvalue and eigenvector problem (Jolliffe & Cadima, 2016; Raj, 2019). A mathematical description of the algorithm can be found in Nsugbe et al. (2020), and the first three PCs were used for the classification problems as they represented 98% of the information from the data.

RPCA: a candidate feature vector matrix that comprises typically of both a low (L) and a sparse (S) component. The RPCA is effectively a matrix separation technique that finds the L and S component of a matrix by finding the solution to an optimization problem termed the Principal Component Pursuit, which capably recovers the sparse representation of a feature vector from a "contaminated" matrix comprising of uncertainties and outliers (Candes et al., 2009; Zhou et al., 2010). In contrast to the traditional PCA that applies the L_2 optimization function, which in turn makes it susceptible to outliers, the RPCA applies the sparsity-based approach during the optimization exercise providing it with a degree of robustness from outliers in comparison (Candes et al., 2009; Zhou et al., 2010).

The key information for a particular application can either be in the L or S component. For the application presented in this paper, it was noted that the S variable maximized the classification accuracy and therein would form the component used for further exercises in this paper when the RPCA is applied. Further mathematical framework of the RPCA can be seen in Candes et al. (2009) and Zhou et al. (2010).

SAC: Autoencoders are composed of artificial neural networks and are primed toward producing a compressed lower dimensional representation of an input feature vector in an unsupervised fashion (Møller, 1993; Ng, 2012; Olshausen & Field, 1997). A SAC consists of an encoder which encodes low dimensional latent properties of the feature vector, and the decoder decodes this latent information into a reconstruction of the original feature vector with an associated reconstructed error (Møller, 1993; Ng, 2012; Olshausen & Field, 1997).

Mathematically, assuming an input feature vector $x \in R^{D_x}$, the encoder assigns x to another vector $z \in R^{D(1)}$, as seen as $z = h^{(1)}\left(W^{(1)}x + b^{(1)}\right)$, where the superscript (1) represents the first layer; $h^{(1)} : R^{D(1)} \to R^{D(1)}$ is an encoder transfer function, $W^{(1)} \in R^{D(1)\,X\,D_x}$ is a weight matrix, and $b^{(1)} \in R^{D(1)}$ represents a bias vector (Møller, 1993; Ng, 2012; Olshausen & Field, 1997). The decoder subsequently maps the now encoded version z into a reconstructed estimate of the input vector x and can be written as $\hat{x} = h^{(2)}\left(W^{(2)}x + b^{(2)}\right)$; here the superscript (2) indicates the second layer of the decoder, $h^{(2)} : R^{D_x} \to R^{D_x}$ represents the decoder's transfer function, $W^{(1)} \in R^{D_x\,X\,D^{(1)}}$ represents a weight matrix, and $b^{(2)} \in R^{D_x}$ is a bias vector. In order for the autoencoder to adopt a sparse representation, a regularization factor is also included and is a function of the average activation value of a neuron, as follows (Møller, 1993; Ng, 2012; Olshausen & Field, 1997):

$$\widehat{\rho}_i = \frac{1}{n}\sum_{j=1}^{n} z_i^{(1)}(x_j) = \frac{1}{n}\sum_{j=1}^{n} h\left(w_i^{(1)T} x_j + b_i^{(1)}\right) \quad (1)$$

where n is the number of training examples, x_j represents the training example, $w_i^{(1)T}$ represents the ith row of the weighting matrix W, and $b_i^{(1)}$ is the ith input of the bias vector $b^{(1)}$. The inclusion of the regularization factor is a means of getting the neurons within the hidden layer to respond to a constrained subset of features that are present within a subset of the training data.

A common regularization term is the Kullback–Leibler (KL) divergence, which measures the difference between distributions. Its mathematical formulation can be seen below:

$$\begin{aligned}\Omega_{\text{sparsity}} &= \sum_{i=1}^{D^{(1)}} KL(\rho||\widehat{\rho}_i) \\ &= \sum_{i=1}^{D^{(1)}} \rho \log\left(\frac{\rho}{\widehat{\rho}_i}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\widehat{\rho}_i}\right)\end{aligned} \quad (2)$$

The aim is typically to minimize the function to zero, thus driving ρ and $\widehat{\rho}_i$ to be equivalent, and effectively grows larger as the values diverge away.

The final cost function for a sparse autoencoder can be expressed as follows:

$$E = \frac{1}{N}\underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K}(x_{kn} - \widehat{x}_{kn})^2}_{\text{Mean Squared Error term}} + \underbrace{\beta * \Omega_{\text{sparsity}}}_{\text{Sparsity Regularisation}} \quad (3)$$

where n is the number of examples, k represents the number of variables within the training data, and β is the coefficient for the sparsity regularization term.

The number of encoded hidden representations within the autoencoder was selected to be three, which was the same number as the PCA dimension that carried 98% of the information from the feature vector, as mentioned earlier.

3.2. Unsupervised learning methods

Gaussian mixture model (GMM): It is a class of mixture models which are a probabilistically driven unsupervised learning approach (Reynolds, 2015; Richardson & Green, 1996). The mathematical architecture for the GMM can be described by the model mean and covariance, and a multidimensional representation of the GMM model can be seen as follows:

$$p(\vec{x}) = \sum_{i=1}^{K} \pi_i N\left(\vec{x}|\vec{\mu_i}, \sum_i\right) \quad (4)$$

$$N\left(x|\mu_i, \sum_i\right) = \frac{1}{\sqrt{(2\pi)^K|\sum_i|}}\exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu_i})^T \sum_i - 1(\vec{x} - \vec{\mu_i})\right) \quad (5)$$

$$\sum_{i=1}^{K} \pi_i = 1 \quad (6)$$

where, for a given data point $\vec{x}$, μ_i represents the mean, $\sum_i$ is the covariance, N represents the assumption of a Gaussian distribution, K is the number of mixture components, while π_i is a normalization component to ensure all probabilities sum up to 1. The cluster forming and learning process is based on the expectation maximization (E-M) process, which is driven by the maximum likelihood estimation sequence (Reynolds, 2015; Richardson & Green, 1996). The hard clustering option was adopted in this work, which meant that data points belonged to a single cluster only.

K-means: It is an unsupervised learning method which separates data points into distinct clusters where it assigns the center of a candidate cluster as a centroid and uses the Euclidean distance metric as an objective function (Dabbura, 2020; Likas et al., 2003). The K-means algorithm uses the E-M framework to assign data points to clusters, where the E step involves $\sum_{i=1}^{m}\sum_{k=1}^{K} argmin_j||x^i - \mu_k||^2$, where x^i is a data point and μ_k is the centroid mean, while the M step involves the recalculation of the class centroid using the expression $\mu_k = \frac{\sum_{i=1}^{m} w_{ik}x^i}{\sum_{i=1}^{m} w_{ik}}$, where w_{ik} is a binary metric used to indicate whether or not a data point belongs in a certain class (Dabbura, 2020; Likas et al., 2003).

Spectral clustering (SC): This clustering method is a graph-based method for the clustering of data points. Its approach to clustering can either involve indicating the number of clusters to the algorithm or by computing the solution to the Laplacian matrix, which provides an indication of the extent that a graph at a particular vertex is similar/different from its neighboring vertices and uses this as a way to find corresponding eigenvectors which indicate how may clusters to split the graph into (Bach & Jordan, n.d.; Ng et al., 2002; von Luxburg, 2007). The steps involved in the SC algorithm include the following for a candidate data point: definition of a local neighborhood followed by the computing of the pairwise distance $Dist_{i,j}$ for all points i and j, application of the kernel transformation to map distance measures to similarity measures using the Kernel transformation $S_{i,j} = \exp\left(-\left(\frac{Dist_{i,j}}{\sigma}\right)^2\right)$, where S is the similarity matrix, σ is the kernel scale factor, optional calculation of the Laplacian matrix, clustering of the points using a K-means cluster assignment approach, and assignment of a further data point to their corresponding clusters (Bach & Jordan, n.d.; Ng et al., 2002; von Luxburg, 2007).

As part of the model tuning process, the number of clusters was specified to be 2, stemming from prior knowledge that the cluster assignment was a 2 class problem, while each model was run for 5 iterations with the maximum value at the end of the iteration cycle selected.

4. Experimental Section and Results

4.1. Cluster validity index

The following performance indexes were chosen to characterize the effectiveness of the clustering methods as follows:

Dunn index: This cluster index is primed toward providing a high score for clusters that are compact with minimal variance within members of the same cluster, with a maximal separation between classes (Ben Ncir et al., 2021; Bezdek & Pal, 1995; Legány et al., 2006). The main limitation of this clustering index is the scaling of computational cost associated with dealing with high dimensional data (Ben Ncir et al., 2021; Bezdek & Pal, 1995; Legány et al., 2006).

Assuming a situation where there are m clusters, the Dunn index can be defined as follows:

$$Dunn's\ Index = \frac{min\ \delta(C_i, C_j)}{\max d(x,y)} \quad (7)$$

where $\delta(C_i, C_j)$ is the inter cluster distance between candidate clusters C_i and C_j, $\max d(x,y)$ is the maximum distance from an n-dimensional vector from which x and y reside and belong to the same cluster.

Accuracy: The accuracy is computed as the ratio between the total number of correctly assigned data points to the total number of points in the cluster expressed as a percentage, as given as follows:

Table 2
The various features, their description, and value range

Raw data	Dunn index	Accuracy %	Autoencoder	Dunn index	Accuracy %	PCA	Dunn index	Accuracy %	RPCA	Dunn index	Accuracy %
GMM	0.172	63	GMM	0.0267	61	GMM	0.0773	59	GMM	0.0773	72
K-means	0.0232	69	K-means	0.0267	69	K-means	0.0232	69	K-means	0.0620	56
SC	0.183	22	SC	0.0232	81	SC	0.1270	59	SC	0.1830	76

$$Accuracy = \frac{no.\ of\ points\ correctly\ assigned}{total\ number\ of\ points\ in\ cluster} * 100 \quad (8)$$

Table 2 shows the results for the various preprocessing methods considered, alongside the candidate unsupervised learning methods. It can be seen that there is not a substantial correlation between the Dunn index and the cluster accuracy, the reason for this is thought to be due to the substantial overlap between the data clusters, therein making the clusters inhomogeneous and nonlinear, with the results implying that the application of the Dunn index is limited for this kind of application. The accuracy results from the raw data showed a slight superiority for the K-means when compared with the GMM, with the SC recording the lowest accuracy and therein showing limited clustering capability in dealing with the raw data without preprocessing, which is likely to contain noise and various sources of uncertainties.

For preprocessed data prior to clustering, for the case of the autoencoder, the SC produces the best accuracy, showing a strongly improved performance from when the raw data were used for clustering–this was followed by the K-means algorithm. In the case of the PCA preprocessing, the K-means produced the best accuracy ahead of both the GMM and SC, while in the case of the RPCA, the SC and GMM produced the best accuracy.

In contrast to the prior study done with supervised learning, that is, the performance of the models shown in Table 2 possess a weaker prediction power when compared with the prior work where accuracies in the range of 90%+ were obtained (Nsugbe et al., Unpublished results). This goes to highlight the superiority of the supervised learning methods when compared with the unsupervised methods in terms of model accuracy. The strength of the unsupervised learning method continues to be the potential of a fully automated workflow which requires minimal intervention and therein can serve as a means of saving time and resources, in comparison to the supervised learning method. However, from the immediate results this appears to be at the direct expense of predictive power, which may reduce its appeal particularly in a clinical setting where the implication of a false prediction could have a combination of health and financial implications.

Part of the challenge in the clustering of this type of data is due to the information emanating from a qualitative source which can be expected to include uncertainties, noise, and outliers, although preprocessing methods have been applied and show potential toward improving model performance, an inhomogeneous cluster overlap continues to exist in the data between classes, as can be seen in Figure 2.

For physiological-based signals, Nsugbe et al. have shown in previous work that the application of an appropriate signal decomposition method can contribute toward increasing cluster separability and overlap and allow for low complexity classifiers to distinguish between classes (Nsugbe, Samuel, et al., 2021; Nsugbe & Sanusi, 2021). The inclusion of decomposition methods in this case is unfeasible due to the information source emanating from a questionnaire and not a signal source, therein

Figure 2
A plot showing a comparison between the actual data cluster projected with PCA and autoencoder SC

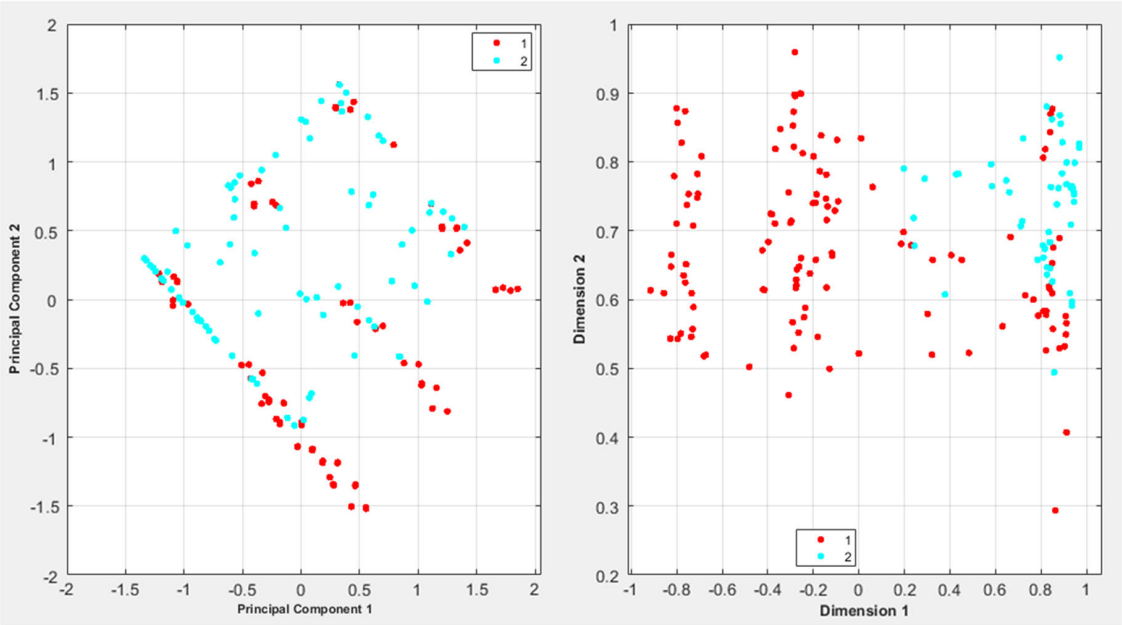


Table 3
Characteristic table showing the various clustering methods used in this study (MathWorks, n.d.)

Clustering method	Basis of algorithm	Requires specified number of clusters	Cluster shape assumption
GMM	Mixture models comprising of Gaussians	Yes	Spheres with different diagonal covariance
K-means	Euclidean distance between centroids, points	Yes	Spheres with equal diagonal covariance
SC	Application of graph theory and distance between graph nodes	Yes/Also capable of determining cluster numbers through Laplacian eigenvalues	Not specified

implying the limitation of unsupervised learning methods and in this particular setting.

Comparing the architectures of unsupervised learning with their supervised counterparts, the unsupervised learning methods used in this study are based around distance proximity in feature space and assumptions around the expected shape of the data clusters themselves. This constrains their overall effectiveness, especially in the case where there exists an inhomogeneous spread of the data clusters, as seen in this work. In the case of the supervised learning methods, aside from receiving labels to pair the data with, these classifiers have a lot of complexity and are able to perform advanced coordinate transformation on the data in a higher dimensional space to promote greater cluster separation, such as the kernel trick in support vector machines, and the hidden layers within a deep neural network (Nsugbe, Obajemu, et al., 2021; Nsugbe, Phillips, et al., 2020). A characteristic table of the various clustering methods used as part of this study can be seen in Table 3:

From the results obtained and the information presented in Table 2, it can be assumed that the strength of the SC algorithm is its ability to not have a specified cluster shape assumption, which in turn allows it to best cluster inhomogeneous data that are nonlinear and carry cluster overlap.

In terms of runtime of the proposed method, once the models have been trained it was seen that predictions could be made within 3–5 s on average with a Core i5 laptop with 4GB RAM.

5. Conclusion and Future Work

The recent decline in male fertility has prompted the need for effective, affordable, and accessible means of semen quality prediction. As an affordable means of semen quality prediction, AI has been used toward predicting semen quality with data from a qualitative questionnaire comprising select factors which are identified by the medical literature to have an influence on semen quality. Prior work has explored the use of supervised learning for pattern recognition and semen quality prediction—despite these results, the reliance of supervised learning on labels prior to model design implies the dependence of the process on an external source of intelligence. The alternative to this is the application of unsupervised learning, which partitions data into various clusters based on an objective function of sorts and is independent of labeling (an external form of intelligence intervention).

In addition to different types of low dimensional embedding methods, three select unsupervised learning algorithms, namely GMM, K-means, and SC, were explored and contrasted as part of this work. Two cluster validation indexes were chosen to help characterize the effectiveness of the algorithms' clusters, namely the Dunn index and the accuracy of the partitioning of the data into the right clusters. The Dunn index proved to be ineffective in this case due to the nature of the data, and thus the effectiveness of the algorithms in question was evaluated using the cluster accuracy. It was seen that the SC algorithm recorded the highest accuracy after preprocessing with the SAC, with the reason for this thought to be due to the arbitrary nature of the cluster shape assumed by the algorithm.

Further work in this area would now involve continued exploration of unsupervised learning methods for further effective classification methods, and in particular, methods whose configuration is similar to that of the SC (Samuel et al., 2020), that is, no cluster shape assumption, to observe the extent which data of this nature can be correctly clustered, and what clustering methods best suit the data source.

Acknowledgments

The authors would like to thank Dr Michael Provost for providing thoughts and feedback on the manuscript, and Brian Kerr from Kerr Editing for proofreading the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Adorf, C. S., Moore, T., Melle, Y., & Glotzer, S. (2019). Analysis of self-assembly pathways with unsupervised machine learning algorithms. *The Journal of Physical Chemistry B*, *124*, 69–78. https://doi.org/10.1021/acs.jpcb.9b09621

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. In M. W. Berry, A. Mohamed, & B. W. Yap (Eds.), *Supervised and unsupervised learning for data science* (pp. 3–21). Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2_1

Andrade-Rocha, F. T. (2013). Temporary impairment of semen quality following recent acute fever. *Annals of Clinical and Laboratory Science*, *43*, 94–97.

Bach, F. R., & Jordan, M. I. (n.d.). *Learning Spectral Clustering* (UCB/CSD-03-1249; p. 14). University of California, Berkeley. Retrieved 24 June 2021, from https://www2.eecs.berkeley.edu/Pubs/TechRpts/2003/CSD-03-1249.pdf

Bali, J., Garg, R., & Bali, R. T. (2019). Artificial intelligence (AI) in healthcare and biomedical research: Why a strong computational/AI bioethics framework is required? *Indian Journal of Ophthalmology*, *67*, 3. https://doi.org/10.4103/ijo.IJO_1292_18

Ben Ncir, C.-E., Hamza, A., & Bouaguel, W. (2021). Parallel and scalable Dunn Index for the validation of big data clusters. *Parallel Computing*, *102*, 102751. https://doi.org/10.1016/j.parco.2021.102751

Bezdek, J., & Pal, N. (1995). *Cluster validation with generalized Dunn's indices* (pp. 190–193). https://doi.org/10.1109/ANNES.1995.499469

Bidgoli, A. A., Komleh, H. E., & Mousavirad, S. J. (2015). Seminal quality prediction using optimized artificial neural network with genetic algorithm. In *2015 9th International Conference on Electrical and Electronics Engineering (ELECO)* (pp. 695–699). https://doi.org/10.1109/ELECO.2015.7394596

Bishop, J. (2002). Views into the Chinese room: New essays on Searle and artificial intelligence. In *Minds and Machines—MIND MACH* (Vol. 15).

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*, 106. https://doi.org/10.1186/1471-2105-14-106

Butler. (2015, May 22). *Underdeveloped: Healthcare in Developing Nations*. Orbis Biosciences. https://orbisbio.com/underdeveloped-healthcare-in-developing-nations/

Candes, E. J., Li, X., Ma, Y., & Wright, J. (2009). Robust principal component analysis? *ArXiv:0912.3599 [Cs, Math]*. http://arxiv.org/abs/0912.3599

Carlsen, E., Andersson, A.-M., Petersen, J. H., & Skakkebaek, N. E. (2003). History of febrile illness and variation in semen quality. *Human Reproduction (Oxford, England)*, *18*, 2089–2092. https://doi.org/10.1093/humrep/deg412

Dabbura, I. (2020, August 10). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Medium. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Encyclopedia Britannica. (n.d.). *Spermatogenesis | Description & Process*. Encyclopedia Britannica. Retrieved 24 June 2021, from https://www.britannica.com/science/spermatogenesis

Gerrits, T., & Shaw, M. (2010). Biomedical infertility care in sub-Saharan Africa: A social science – review of current practices, experiences and view points. *Facts, Views & Vision in ObGyn*, *2*, 194–207.

Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J., & Johnsson, M. (2012). Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, *39*, 12564–12573. https://doi.org/10.1016/j.eswa.2012.05.028

Grant, J., Hoorens, S., Sivadasan, S., Loo, M. V. H., Davanzo, J., Hale, L., & Butz, W. (2006). Trends in European fertility: Should Europe try to increase its fertility rate . . . or just manage the consequences? *International Journal of Andrology*, *29*, 17–24. https://doi.org/10.1111/j.1365-2605.2005.00634.x

Gunes, S., Hekim, G. N. T., Arslan, M. A., & Asci, R. (2016). Effects of aging on the male reproductive system. *Journal of Assisted Reproduction and Genetics*, *33*, 441–454. https://doi.org/10.1007/s10815-016-0663-y

Hadi, H. A., Hill, J. A., & Castillo, R. A. (1987). Alcohol and reproductive function: A review. *Obstetrical & Gynecological Survey*, *42*, 69–74.

IBM Cloud Education. (2021, May 7). *What is Supervised Learning?* [IBM]. https://www.ibm.com/cloud/learn/supervised-learning

Inhorn, M. C. (2003). Global infertility and the globalization of new reproductive technologies: Illustrations from Egypt. *Social Science & Medicine (1982)*, *56*, 1837–1851. https://doi.org/10.1016/s0277-9536(02)00208-3

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*, 20150202. https://doi.org/10.1098/rsta.2015.0202

Jørgensen, N., Andersen, A.-G., Eustache, F., Irvine, D. S., Suominen, J., Petersen, J. H., . . . Skakkebæk, N. E. (2001). Regional differences in semen quality in Europe. *Human Reproduction*, *16*, 1012–1019. https://doi.org/10.1093/humrep/16.5.1012

Jung, A., Schuppe, H. C., & Schill, W. B. (2001). [Fever as etiology of temporary infertility in the man]. *Der Hautarzt; Zeitschrift Fur Dermatologie, Venerologie, Und Verwandte Gebiete*, *52*, 1090–1093. https://doi.org/10.1007/s001050170018

Kerns, J. (2017, February 15). *What's the difference between weak and* strong *AI?* Machine Design. https://www.machinedesign.com/markets/robotics/article/21835139/whats-the-difference-between-weak-and-strong-ai

Kotsiantis, S. (2007). Supervised Machine learning: A review of classification techniques. *Informatica (Ljubljana), 31.*

Legány, C., Juhász, S., & Babos, A. (2006). Cluster validity measurement techniques. In *Proceedings of the 5th WSEAS international conference on artificial intelligence, knowledge engineering and data bases* (pp. 388–393).

Levine, H., Jørgensen, N., Martino-Andrade, A., Mendiola, J., Weksler-Derri, D., Mindlis, I., . . . Swan, S. H. (2017). Temporal trends in sperm count: A systematic review and meta-regression analysis. *Human Reproduction Update*, *23*, 646–659. https://doi.org/10.1093/humupd/dmx022

Levine, R. J. (1999). Seasonal variation of semen quality and fertility. *Scandinavian Journal of Work, Environment & Health*, *25*(Suppl. 1), 34–37; discussion 76–78. https://pubmed.ncbi.nlm.nih.gov/10235406/

Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*, 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

MathWorks. (n.d.). *Choose Cluster Analysis Method—MATLAB & Simulink—MathWorks United Kingdom* [MathWorks]. Retrieved 24 June 2021, from https://uk.mathworks.com/help/stats/choose-cluster-analysis-method.html

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, *6*, 525–533. https://doi.org/10.1016/S0893-6080(05)80056-5

Neto, F. T. L., Bach, P. V., Najari, B. B., Li, P. S., & Goldstein, M. (2016). Spermatogenesis in humans and its affecting factors. *Seminars in Cell & Developmental Biology*, *59*, 10–26. https://doi.org/10.1016/j.semcdb.2016.04.009

Ng, A. (2012). CS229 Lecture Notes. Akademik.Bahcesehir.Edu.Tr. ~https://akademik.bahcesehir.edu.tr/~tevfik/courses/cmp5101/cs229-notes1.pdf

Ng, A., Jordan, M., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems, *14*. http://papers.neurips.cc/paper/2092-on-spectral-clustering-analysis-and-an-algorithm.pdf

Nsugbe, E., Obajemu, O., Samuel, O. W., & Sanusi, I. (2021). Application of noninvasive magnetomyography in labour imminency prediction for term and preterm pregnancies and ethnicity specific labour prediction. *Machine Learning with Applications*, *5*, 100066. https://doi.org/10.1016/j.mlwa.2021.100066

Nsugbe, E., Phillips, C., Fraser, M., & McIntosh, J. (2020). Gesture recognition for transhumeral prosthesis control using EMG and NIR. *IET Cyber-Systems and Robotics*, *2*, 122–131. https://doi.org/10.1049/iet-csr.2020.0008

Nsugbe, E., Samuel, O. W., Asogbon, M. G., & Li, G. (2020). A self-learning and adaptive control scheme for phantom prosthesis control using combined neuromuscular and

brain-wave bio-signals. *Engineering Proceedings*, *2*, 59. https://doi.org/10.3390/ecsa-7-08169

Nsugbe, E., Samuel, O. W., Asogbon, M. G., & Li, G. (2021). Contrast of multi-resolution analysis approach to transhumeral phantom motion decoding. *CAAI Transactions on Intelligence Technology*, https://doi.org/10.1049/cit2.12039

Nsugbe, E., Samuel, O. W., & Sanusi, I. (Unpublished results). *A cybernetic system towards an affordable male fertility prediction using qualitative lifestyle and environmental factors information ensemble*.

Nsugbe, E., & Sanusi, I. (2021). *Towards an affordable magnetomyography instrumentation and low model complexity approach for labour imminency prediction using a novel multiresolution analysis* [Preprint]. *Preprints*. https://doi.org/10.22541/au.161289481.19912239/v1

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*, 3311–3325. https://doi.org/10.1016/s0042-6989(97)00169-7

Provost, C. (2014, January 29). *One in four young people in developing countries unable to read, says UN*. The Guardian. http://www.theguardian.com/global-development/2014/jan/29/illiteracy-education-young-people-developing-countries

Raj, J. T. (2019, March 14). *Dimensionality Reduction for Machine Learning*. Medium. https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e

Reynolds, D. (2015). Gaussian Mixture Models. In S. Z. Li & A. K. Jain (Eds.), *Encyclopedia of biometrics* (pp. 827–832). Springer US. https://doi.org/10.1007/978-1-4899-7488-4_196

Richardson, S., & Green, P. (1996). On Bayesian analysis of mixture with unknown number of components. *Journal of the Royal Statistical Society: Series B*, *60*. https://doi.org/10.2307/2985194

Samuel, O. W., Yang, B., Geng, Y., Asogbon, M. G., Pirbhulal, S., Mzurikwao, D., … Li, G. (2020). A new technique for the prediction of heart failure risk driven by hierarchical neighborhood component-based learning and adaptive multi-layer networks. *Future Generation Computer Systems*, *110*, 781–794. https://doi.org/10.1016/j.future.2019.10.034

Sergerie, M., Mieusset, R., Croute, F., Daudin, M., & Bujan, L. (2007). High risk of temporary alteration of semen parameters after recent acute febrile illness. *Fertility and Sterility*, *88*, 970.e1-7. https://doi.org/10.1016/j.fertnstert.2006.12.045

Sharma, R., Agarwal, A., Rohra, V. K., Assidi, M., Abu-Elmagd, M., & Turki, R. F. (2015). Effects of increased paternal age on sperm quality, reproductive outcome and associated epigenetic risks to offspring. *Reproductive Biology and Endocrinology: RB&E*, *13*, 35. https://doi.org/10.1186/s12958-015-0028-x

Sharpe, R. M. (2010). Environmental/lifestyle effects on spermatogenesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*, 1697–1712. https://doi.org/10.1098/rstb.2009.0206

S. Brunton. (2020, January 23). *Matrix Completion and the Netflix Prize*. https://www.youtube.com/watch?v=sooj-_bXWgk

Techopedia. (n.d.). *What is Weak Artificial Intelligence (Weak AI)? – Definition from Techopedia*. Techopedia.Com. Retrieved 24 June 2021, from http://www.techopedia.com/definition/31621/weak-artificial-intelligence-weak-ai

Tomlinson, M. J. (2016). Uncertainty of measurement and clinical value of semen analysis: Has standardisation through professional guidelines helped or hindered progress? *Andrology*, *4*, 763–770. https://doi.org/10.1111/andr.12209

UCI Machine Learning Repository: Fertility Data Set. (n.d.). Uci.Edu. Retrieved 6 June 2021, from https://archive.ics.uci.edu/ml/datasets/Fertility

Verner, D. (2005). What factors influence world literacy? Is Africa different? In *Policy Research Working Paper Series* (No. 3496; Policy Research Working Paper Series). The World Bank. https://ideas.repec.org/p/wbk/wbrwps/3496.html

Vignera, S., Condorelli, R., Balercia, G., Vicari, E., & Calogero, A. (2012). Does alcohol have any effect on male reproductive function? A review of literature. *Asian Journal of Andrology*, *15*. https://doi.org/10.1038/aja.2012.118

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*, 395–416. https://doi.org/10.1007/s11222-007-9033-z

Your Fertility. (n.d.). *Understanding how to improve your chance of having a baby*. Your Fertility. https://www.yourfertility.org.au/everyone/age#:~:text=Age%20and%20sperm,healthy%20sperm%20than%20younger%20men

Zhou, Z., Li, X., Wright, J., Candes, E., & Ma, Y. (2010). Stable Principal Component Pursuit. *ArXiv:1001.2363 [Cs, Math]*. http://arxiv.org/abs/1001.2363

How to Cite: Nsugbe, E. (2023). Toward a Self-Supervised Architecture for Semen Quality Prediction Using Environmental and Lifestyle Factors. *Artificial Intelligence and Applications* 1(1), 35–42, https://doi.org/10.47852/bonviewAIA2202303