Proceedings of the 18th Conference on Computational Intelligence Methods for Bioinformatics & Biostatistics (CIBB 2023)

Computational Intelligence using a Supervised Autoencoder Neural Network with multi-categorical medical data for Thyroid Nodule Diagnosis.

Cyprien Gille[1],Grégoire D'Andrea[2,3], Thierry Pourcher[3], Michel Barlaud[1]*

[1] Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S), Université Côte d'Azur (UCA), Centre National de la Recherche Scientifique (CNRS), Sophia Antipolis, France.
[2] Otorhinolaryngology and Head and Neck surgery department. Nice University Hospital, Université Côte d'Azur, Nice, France.
[3] Université Côte d'Azur, Commissariat à l'Energie Atomique et aux énergies alternatives, Institut Frédéric Joliot, Service Hospitalier Frédéric Joliot, Transporters in Imaging and Radiotherapy in Oncology (TIRO) laboratory, Nice, France.

*corresponding author

Keywords: *Computational Intelligence, Supervised Autoencoder, Cancer Diagnosis, Confidence Score, Graphical User Interface.*

Abstract. Thyroid Nodules are more and more widespread, with the frequency of thyroid cancers increasing accordingly. In routine practice, the clinician's goal is to determine if a nodule is malignant or suspicious, and thus requires surgery. However, nearly 33% of thyroid nodules fall into the 'indeterminate' classification segment, often leading to inappropriate surgery.
In this paper, we propose an accurate method to predict the malignancy of thyroid nodules using a categorical thyroid nodules dataset. The diagnosis is achieved through computational intelligence using a new supervised autoencoder neural network. A proof-of-concept Graphical User Interface was developed to demonstrate the medical usefulness of the latent space of the trained network.
Results show that the Supervised auto encoder Neural Network provides a diagnosis with a high accuracy of 85% and outperforms classical classification methods on all evaluation metrics. Aside from the accuracy itself, our computational intelligence method provides three medically interesting benefits: i) a prediction confidence score and consistent feature ranking, ii) an interpretative latent space, and iii) a flexibility for controlling errors of type I or errors of type II. All of those benefits are valuable from the clinician's point of view.

1 Study objectives

1.1 Medical requirements

Thyroid nodules (TN) are widespread in the human population, and their frequency is expected to continue to increase [1, 2, 3, 4]. In routine practice, the aim is to identify benign nodules that may benefit from surveillance, and malignant or suspicious nodules for which surgery may be required. However, despite optimal diagnostic management, approximately 33 $\%$ of all TNs fall into the grey area of the Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) cytological classification, where the diagnosis of benignity or malignancy remains uncertain. Thus, there is an urgent requirement to find an accurate method to improve the diagnosis of the nature of Indeterminate Thyroid Nodules (ITN), in order to reduce the number of inappropriate surgical procedures.

1.2 Dataset

In this study, we use a database compiled by Dr. G. D'Andréa, comprised of mostly categorical clinical and cytological data [5]. It contains data ranging from 2010 to 2020, originating from four french hospitals. We use three sources of nodule information: clinical, echography and cytology data. The in-house dataset used in this study contains 1337 nodules from 1290 patients: Nodules are frequently multiple, and do not all share the same ultrasonographic and cytological characteristics within a patient - thus we want to predict the risk of malignancy for a nodule, not a patient. Each nodule is described through 159 mostly categorical features (clinical, from an echography, from a cytology, from surgery or from an extemporaneous biopsy...) [5]. Of the initial features, we picked 12 features of higher medical importance reported in the literature [1, 2, 3, 4]: the TI-RADS score, the cytological category according to TBSRTC, the cytological sub-classification according to Valderrabano et al. [4], the nodule's size in echography, the presence of colloid in the cytologic sample, clinical hyperthyroidism, the Hashimoto's thyroiditis, the patient's age and gender, clinical hypothyroidism, the existence of a thyroidopathy of any kind, and the cellularity of the cytologic sample.

All of these features are categorical, with two to six possible values, except for the age and the echography size. For these two numerical features, we use categorical bins with medically sound ranges. In our dataset, 144 nodules had a missing value for at least one feature. We addressed this classical problem by filling those values with the mean over all nodules for the corresponding feature.

2 Computational Intelligence using a Supervised Autoencoder Neural Network

Neural Networks have proven their effectiveness in bioinformatics for classification and feature selection [7, 8, 9]. Classical stacked autoencoders [10] were used recently in metabolomic studies. Their biggest strength is their ability to learn a representation of the data, typically in a space of lower dimension than the input space. As such, they are often used for dimensionality reduction [11], and have applications in the medical field as data denoisers and relevant feature selectors [10]. A widely used type of autoencoders is the Variational Autoencoder (VAE). The VAE adds the assumption that the encoded data follows a prior, a gaussian distribution, and thus combines the reconstruction loss with a distance function (between the gaussian prior and the actual learned distribution) [12].

2.1 Supervised Autoencoder Neural Network

In this study, we use a supervised autoencoder (SAE) neural network, introduced in 2021 by Barlaud and Guyard in [13], where we relax the parametric distribution assumption present in classical VAEs. In its place, we use labels present in the data (in our case, the malignancy of a nodule) and a classification loss to encourage the latent space to fit the actual distribution of the data. See [13, 14]. Figure 1 depicts the main constituent blocks of our proposed approach. Note that we added a "soft max" block to our autoencoder to compute the classification loss.

Let X be the matrix ($m \times d$) of the dataset (m nodules, d features), and Y the labels in $\{0, \dots, k\}$, with k the number of classes. Let Z be the encoded latent vectors, $\widehat{X}$ the reconstructed data and W the weights of the neural network. Note that the dimension of the latent space k corresponds to the number of classes. In this study, $k = 2$.

We compute the weights W of the autoencoder by minimizing the following criterion :

$$Loss(W) = \mathcal{C}_{\lambda_1,\lambda_2}(Z, Y) + \psi(\widehat{X} - X) \text{ s.t. } \|W\|_1^1 \leq \rho. \tag{1}$$

$$\mathcal{C}_{\lambda_1,\lambda_2}(Z, Y) = \begin{cases} \lambda_1 \mathcal{C}(Z, Y) & if\, Y = 0 \\ \lambda_2 \mathcal{C}(Z, Y) & if\, Y = 1 \end{cases} \tag{2}$$

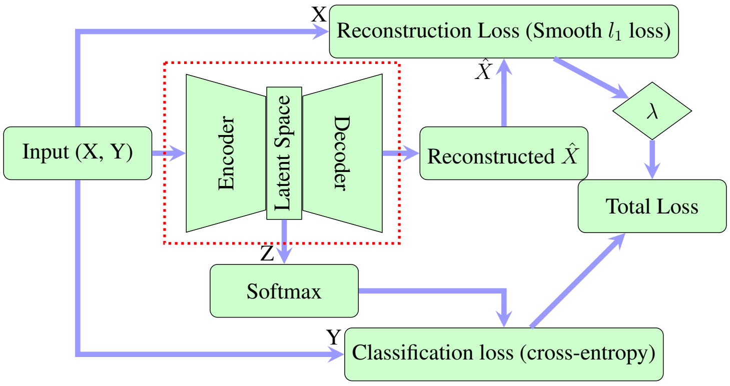


Figure 1: Supervised Autoencoder Neural Network Architecture.

Where λ_1 and λ_2 are the weights of errors of type I and type II respectively and ρ a parameter which controls the sparsity. We use the weighted Cross Entropy Loss for the classification loss $\mathcal{C}$. We use the robust Smooth ℓ_1 (Huber) Loss as the reconstruction loss ψ.

2.2 Statistical evaluation

The supervised autoencoder and its training and evaluation routines were programmed using the PyTorch package. We used the NetBIO SAE, a linear fully connected network, which has an input layer of $d = 12$ neurons, 1 hidden layer of 96 neurons followed by a ReLU activation function, and a latent layer of dimension 2 (the number of classes). The decoder is symmetrical to the encoder. Since the original data is heterogeneous, we scale the data to zero mean and unit variance. We train the supervised autoencoder following a 4-fold cross validation process, as well as several randomness seeds. Table 1 shows that the model obtains competitive-to-superior performance metrics compared to the classical methods (Partial Least Squares Discriminant Analysis, LASSO, and Random Forests). As a reminder, the F1 score is the harmonic mean of precision and recall, incorporating information about both the false positives and negatives; and the AUC is the area under the ROC curve, representing the discriminative power of a classification model. The autoencoder outperforms all of the classical methods in terms of accuracy and F1 score by at least 2.4 %. Note that all standard deviations are very similar.

Optimisation of false-positive versus false-negative detection is often in a trade-off. From a medical point of view, false positives will have to undergo additional tests or even an unnecessary surgery, while false negatives will see their treatment delayed, which can have a severe impact on their health. Thus a false negative test poses greater risks for the patient, and most testing applications are set up to minimise the occurrence of false-negative results. Note that the SAE has the best performances for the false negatives detection. As mentioned earlier, the SAE also allows us to weigh more severely false negatives than false positives using the λ_1 versus λ_2 classification loss weights. Table 1 shows that this SAE-FN variant provides a significant decrease of the false negatives.

Using the DeepLIFT method, implemented in the captum python package, we can determine feature importances for the supervised autoencoder. Table 2 shows the feature rankings, and displays the second advantage of using the supervised autoencoder over classical methods: Features are selected more strongly, i.e. the SAE needs less features to make its prediction: the importances already become relatively negligible for the fifth feature (hyperthyroidism at 0.02). Features are also ranked very consistently across runs; both of these qualities help with interpretability, which is crucial for clinical use.

Method	**PLSDA**	**LASSO**	**RF**	**SAE**	**SAE-FN**
Accuracy	83.24 ± 1.46	83.11 ± 1.36	83.30 ± 1.62	**85.19** ± 1.50	84.81 ±1.46
AUC	82.25 ± 1.74	82.15 ± 1.69	82.36 ± 2.18	**82.99** ± 1.73	81.49 ±1.90
F1 score	81.38 ± 1.72	81.21 ± 1.67	81.20 ± 2.03	**84.02** ± 1.66	82.62 ±1.56
% of FP	**4.79**	5.08	6.89	5.31	9.05
% of FN	11.08	11.08	13.47	10.17	**7.48**

Table 1: Mean Autoencoder performance metrics over 3 seeds and 4 folds of cross validation. Comparison with classical classification methods.

Feature	**PLSDA**	**LASSO**	**RF**	**SAE**
TI-RADS	1	0.85	0.65	**0.97**
Cyto-Bethesda	0.88	1	1	**0.73**
Cyto-subclassif	0.34	0.33	0.42	**0.44**
Echo Size	0.14	0.14	0.18	**0.12**
Age	0.20	0.17	Colloid: 0.05	**Hyperthyroidism: 0.02**

Table 2: Feature rankings by the autoencoder (Means over 3 seeds and 4 folds). Comparison with classical methods.

3 Prognosis in the latent space of the supervised autoencoder with a confidence score

The third and main advantage of the autoencoder is the projection of the data in the latent space by the encoder: as our latent space is of dimension 2 (two target classes), it can very easily be represented, and gives accurate and visually explained predictions. As we can see in Fig. 2, the two classes are distinctly clustered in the latent space. Note that among classical methods, only the PLSDA offers any projection capabilities, and even then, does not cluster data in any meaningful way [14].

The softmax used on the latent space (see Fig 1) gives us a confidence score on the predicted label. This characteristic is only offered by the supervised autoencoder: none of the classical methods provide a sense of how confident the prediction was.

To leverage these advantages, a user-friendly graphical user interface (GUI) was developed, enabling physicians to visualize the impact of each feature on the prediction. A screenshot of it can be seen in figure 2. By modifying input data and observing real-time changes in projection and classification, clinicians can gain insight into the network's confidence and reasoning behind predictions. The code for reproducing all results is made freely available on Github[1].

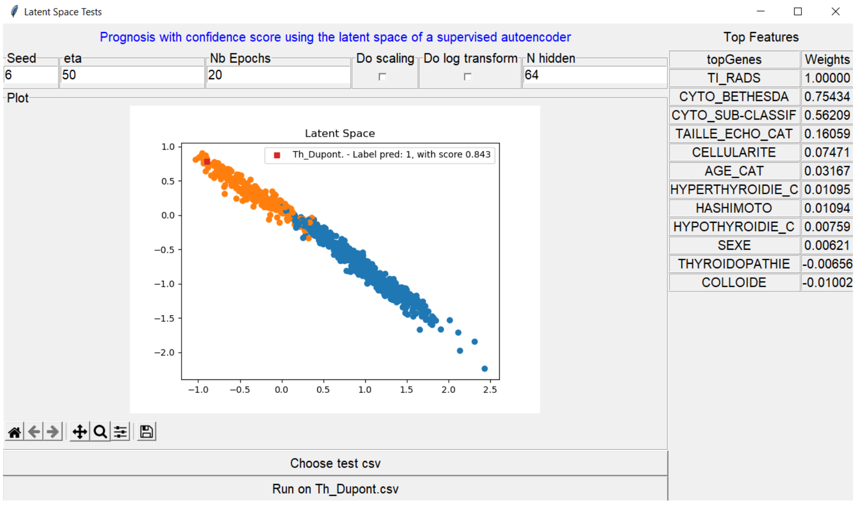
In this example, we can see that the test patient has been classified in class 1 with a confidence score of 0.843.

4 Discussion and conclusion

Beyond its performances, our computational intelligence method using a supervised autoencoder allowed us to create a graphical representation of the malignity prediction of one TN, as a projection of the data into a latent space : having only two dimensions, this latent space makes it easy for clinicians to represent, visualize, and interpret the prediction. Our model also gives a confidence score on the predicted label, which allows estimations of the degree of reliability of the prediction. The proposed method was medically validated by a coauthor clinician at the Otorhinolaryngology and Head and Neck surgery department, Nice University Hospital.

In addition, it is now freely available on Github under the MIT License, which avoids any extra

[1] https://github.com/CyprienGille/Diagnostic-SAE-GUI



Prognosis with confidence score using the latent space of a supervised autoencoder

Seed: 6
eta: 50
Nb Epochs: 20
Do scaling: □
Do log transform: □
N hidden: 64

🟥 Th_Dupont. - Label pred: 1, with score 0.843

Run on Th_Dupont.csv

Top Features

topGenes	Weights
TI_RADS	1.00000
CYTO_BETHESDA	0.75434
CYTO_SUB-CLASSIF	0.56209
TAILLE_ECHO_CAT	0.16059
CELLULARITE	0.07471
AGE_CAT	0.03167
HYPERTHYROIDIE_C	0.01095
HASHIMOTO	0.01094
HYPOTHYROIDIE_C	0.00759
SEXE	0.00621
THYROIDOPATHIE	-0.00656
COLLOIDE	-0.01002

Figure 2: Graphical User Interface for feature testing.

cost related to the use of these tools, thereby seeking to reduce the medico-economic impact of the management of ITNs. To further improve the accuracy of our computational intelligence method in the future, we propose to add metabolomic data [6].

In this study, we showed that using a supervised autoencoder on categorical medical data produced a prognosis on the malignancy of thyroid nodules that was perfectly suited for a physician's use. The autoencoder offered features that were not available with classical classification methods, and that are greatly relevant from the medical practitioner's point of view: strong feature selection, output of a confidence score, and a visually meaningful 2-dimensional latent space. Finally, the flexibility of the SAE to weigh more severely false negatives than false positives allows the clinician to control the false negative rate. A medical perspective would be to keep two SAEs, trained minimising either the FPR or the FNR. The practician would then run the SAE twice for the same patient: If there is agreement on the classification of the nodule by the two networks, then the malignancy risk prediction will be very trustworthy.

Ethics approval

The authors confirm that this retrospective study has been approved by the institutional ethics committees of the University Hospital of Nice and the University Hospital of Toulouse.
In addition, this study met the requirements of the French reference method 04 of the ”*commission nationale de l'informatique et des libertés*” and was declared under the numbers 2204177v0, and F20210608133203. https://www.health-data-hub.fr/projets/amelioration-du-diagnostic-et-de-la-prise-en-charge-chirurgicale-des-nodules-thyroidiens.

The authors declare that they have no competing interests.

Availability of the data and code.

The datasets generated and/or analysed during the present study are available in the Diagnostic-SAE-GUI repository, at
`https://github.com/CyprienGille/Diagnostic-SAE-GUI/tree/main/data`.

References

[1] Bongiovanni M, Bellevicine C, Troncone G, Sykiotis GP. Approach to cytological indeterminate thyroid nodules. Gland Surg. 2019; 8: S98-S104.

[2] Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid. 2016 Jan; 26(1): 1-133.

[3] Cibas ES and Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid. 2017; 27: 1341-6.

[4] Valderrabano P, Khazai L, Thompson ZJ, Sharpe SC, Tarasova VD, Otto KJ, et al. Cancer Risk Associated with Nuclear Atypia in Cytologically Indeterminate Thyroid Nodules: A Systematic Review and Meta-Analysis. Thyroid. 2018; 28: 210-9.

[5] D'Andréa G. Apport de l'intelligence artificielle dans la prise en charge des nodules thyroïdiens indéterminés. Thèse d'exercice de médecine. Faculté de Médecine de Nice (Octobre 2021).

[6] Kell DB. Metabolomics and systems biology: making sense of the soup. Curr Opin Microbiol. 2004 Jun;7(3):296-307.

[7] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017 Sep 1;18(5):851-69.

[8] Leclercq M, Vittrant B, Martin-Magniette ML, Scott Boyer MP, Perin O, Bergeron A, et al. Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. Front Genet. 2019 May 16;10:452.

[9] Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ. Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 2017 Jan;21(1):4-21.

[10] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. J Mach Learn Res. 2010; 11: 3371-3408.

[11] Hinton GE, Zemel RS. Autoencoders, minimum description length and Helmholtz free energy. Advances in neural information processing systems. 1994; 6: 3-10.

[12] Kingma D, Welling M. Auto-encoding variational bayes. In: International Conference on Learning Representation (2014).

[13] Barlaud M and Guyard F. Learning a sparse generative non-parametric supervised autoencoder, In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, TORONTO, Canada, (June 2021).

[14] Chardin D, Gille C , Humbert O, Pourcher T, Barlaud M. Fellow IEEE.: Efficient Diagnosis with confidence score using a new Supervised Autoencoder with structural constraints for clinical metabolomic studies. BMC Bioinformatics. 2022 September 23;361.