

# Few-Shot Learning for Predicting Genetic Biomarkers in Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL)

Ali Aguerd<sup>1\*</sup>, Oumaima Anachad<sup>1</sup>, Asmae Taheri<sup>1</sup>, Faïza Bennis<sup>1</sup> and Fatima Chegdani<sup>1</sup>

<sup>1</sup>Laboratory of Integrative Biology, Faculty of Science Ain Chock, Casablanca, University Hassan II, Morocco

## Abstract:

Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL) is a rare hereditary cerebral small-vessel disorder with an estimated prevalence of 4.6 per 100,000 adults, primarily caused by *NOTCH3* mutations. Its rarity has limited the availability of genetic data, which is important to understand this pathology. Traditional prediction methods require large datasets and fail with limited data. Given these challenges, our study aims to enrich the genetic data on CADASIL. To achieve this, we applied a Few-Shot Learning (FSL) strategy. A total of 4 previously validated CADASIL single nucleotide polymorphisms (SNPs) and 938,544 negative SNPs were extracted from the GWAS catalogue, with their genetic annotations. Based on the assumption of genetic proximity, we generated for each SNP a genomic context string. These strings were embedded into dense vector representations using paraphrase-MiniLM-L6-v2. Similarity scores then ranked candidate SNPs, and the top 100 were identified as novel biomarkers. This *in silico* framework predicted 100 SNPs and 24 genes. It provides potential biomarkers for early diagnosis, insights into disease mechanisms, and candidate therapeutic targets. This study also validates the compatibility of FSL in the context of rare diseases, paving the way for other applications.

**Keywords:** CADASIL, Genetic Biomarkers, Few-Shot Learning, Genomic Embeddings, Genome-Wide Association Studies

---

\* Corresponding author: [aliagu97ninety@gmail.com](mailto:aliagu97ninety@gmail.com)

## 1. Introduction

CADASIL is a chronic disease that affects the small cerebral vessels. This disease is characterized mainly by recurrent subcortical ischemic strokes, a gradual decline in cognitive function, and other neurological symptoms [1]. This disorder has a prevalence of 1.3 - 4.1 cases per 100,000 adults [1]. Taken together, these data confirm the rarity and severity of CADASIL. Indeed, genetic data on this disease are very limited, and most of them concern the *NOTCH3* gene [2]. This lack of data prevents a thorough understanding of this chronic condition, which is therefore incurable [2].

Faced with this challenge, genome-wide association studies (GWAS) found associations between single nucleotide polymorphisms (SNPs) and CADASIL [3]. Machine learning (ML) was also used to predict genetic data [4]. Unfortunately, the lack of genetic data on CADASIL persists, highlighting the limitations of current methodological approaches. GWAS and ML are major approaches, but their performance is reduced when faced with rare diseases with small samples and limited data. Indeed, a methodological approach adapted to rare diseases is required.

Few-Shot Learning (FSL) is a branch of ML that is independent of large data sets. The principle behind this technique is mainly based on the use of a model pre-trained with a large volume of data and its application to new situations [5]. FSL is well suited to the context of rare diseases, as it can predict large amounts of genetic data from a very small sample.

Finally, this FSL-based study aims to expand genetic knowledge about CADASIL. It seeks to predict biomarkers. This genetic data can be used in several ways, such as early diagnosis, the construction of cell signaling pathways, and the development of targeted therapeutic strategies. In addition, this study serves to validate the compatibility of FSL with rare diseases, enabling future extensions to other rare cases.

## 2. Methods

### 2.1. Theoretical Background

GWAS and Linkage Disequilibrium (LD) are the key concepts of our study.

- **GWAS** [6] : This approach is the primary source of our sample. Its principle is based on the comparison of the characteristic SNPs of two groups of individuals, those

with a disease and those who are healthy, which makes it possible to identify certain SNPs that differ significantly between the two groups and finally correlate them with the disease being studied.

- **LD** [7] : This concept forms the basis of our predictive reasoning. Indeed, SNPs that are genetically close have linkage disequilibrium, may have the same function, and therefore be linked to the same pathologies.

### 2.2. Genetic Data

Positive cases comprised 4 CADASIL-Associated SNPs obtained from the GWAS Catalog [8] in TSV format. For the negative cases, we selected a sample of 5,000 SNPs from a pool of 938,544 non-CADASIL SNPs collected from the GWAS catalog. We used a fixed random state (`random_state=42`) for this sampling, to keep the process computationally manageable, while still ensuring a genetically diverse control set. For each SNP, genomic features including its chromosome, precise chromosomal position and its mapped gene(s) were extracted. This collected data constitutes the source of learning for the model used.

### 2.3. Few-shot learning prediction

#### 2.3.1. Genomic Context Encoding

The few-shot learning model requires a prediction criterion, so we constructed a genomic context for each of our SNPs, whether positive or negative.

This genomic context is composed of the carrier chromosome, the exact chromosomal position, and the carrier gene(s) (name, number, and presence of genes characteristic of CADASIL: *NOTCH3* and *HTRA1*).

#### 2.3.2. Vector Representation

In FSL mode, each SNP's genomic context was converted to a textual descriptor following the format:

```
"chr{chromosome}_pos{position}_genes{gene_list}"
```

These descriptors were encoded into dense vector representations using the pre-trained SentenceTransformer model 'paraphrase-MiniLM-L6-v2' [9]. This model was chosen for its speed and simplicity and its compatibility with static data, which is the case with genomic position data. Indeed, this model transforms genetically identical or similar positions into identical or similar vector representations. Consequently, genetically similar SNPs are transformed into similar vector representations, allowing them to be correlated

during the comparison phase. Thus, this model is compatible with static positional data from SNPs and is independent of their functioning and dynamics. In addition, it is suitable for any available hardware resources. The first 20 dimensions of the resulting 384-dimensional embeddings were retained for similarity calculation. This number of embedding dimensions represents a compromise between computational resources, efficiency, and performance.

### 2.3.3. Feature Filtering and Weighted Similarity Calculation

A biologically-informed weighting scheme was applied to filter and prioritize features for similarity computation. The weights were assigned as follows:

- Chromosomal position: 1.0
- Gene count: 0.5
- Presence of CADASIL genes: 3.0
- Embedding dimensions (embed\_0 to embed\_19): 0.8 each

All features underwent RobustScaler normalization before similarity computation. A prototype vector was constructed by computing the weighted mean of

positive CADASIL examples, followed by L2 normalization. Cosine similarity was calculated between this prototype and all negative examples using the weighted feature space. The resulting similarity scores were transformed using a non-linear scaling function (power of 0.8) and clipped to the [0,1] range.

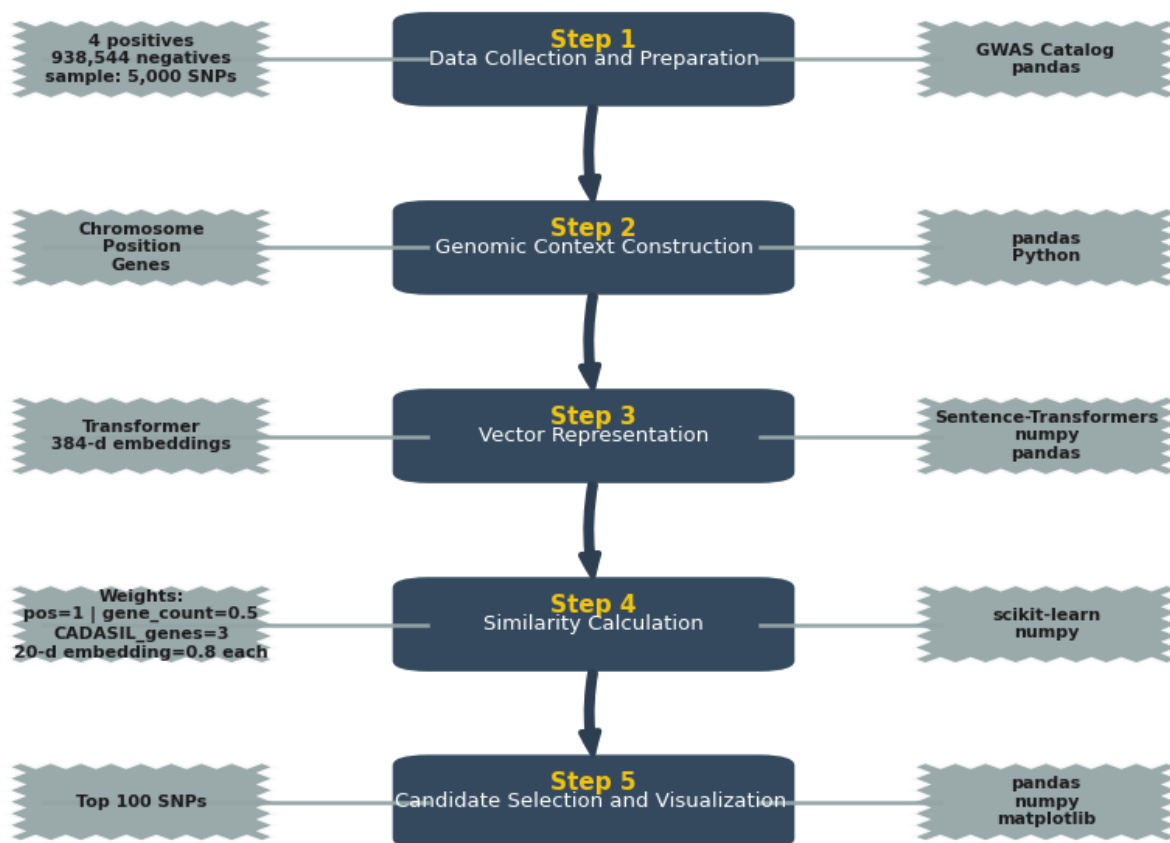
### 2.3.4. Candidate Selection

SNPs were ranked by their similarity scores, with the top 100 candidates selected for further analysis. Model performance was evaluated through distribution analysis of similarity percentiles between positives and predicted SNPs.

### 2.3.5. Analysis environment

We performed all of these steps using Python 3.12.6. The data preprocessing was handled with the pandas library, while numerical computations relied on numpy. For model training, preprocessing, and evaluation, we used scikit-learn. Finally, we created the performance metric visualizations with matplotlib and seaborn.

Figure 1 presents an overview of the different methodological phases:

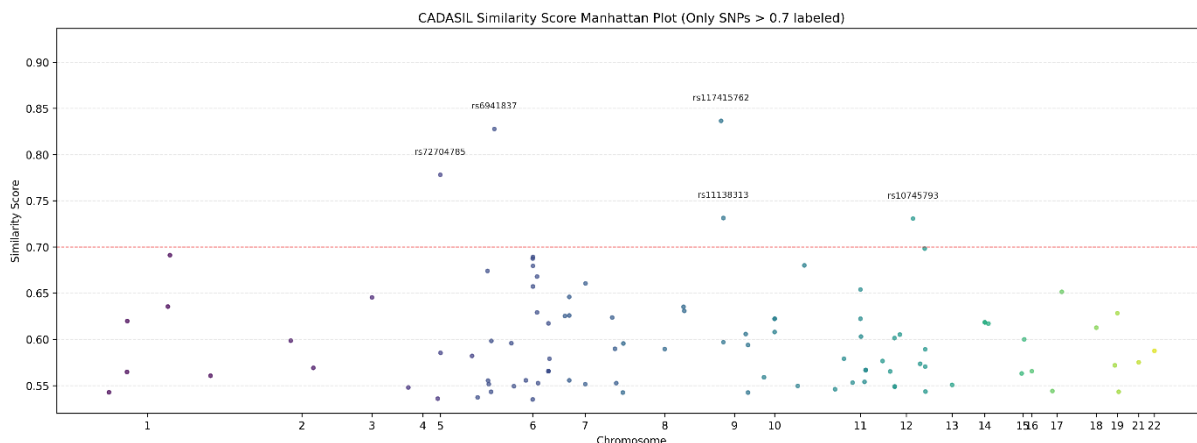


**Figure 1: Few-Shot Learning Workflow for CADASIL SNPs Prediction.** The central boxes present the methodological phases, and the branches show the technical details and tools used

### 3. Results

Figure 2 presents the main result of this study. It shows the top 100 predicted SNPs and their

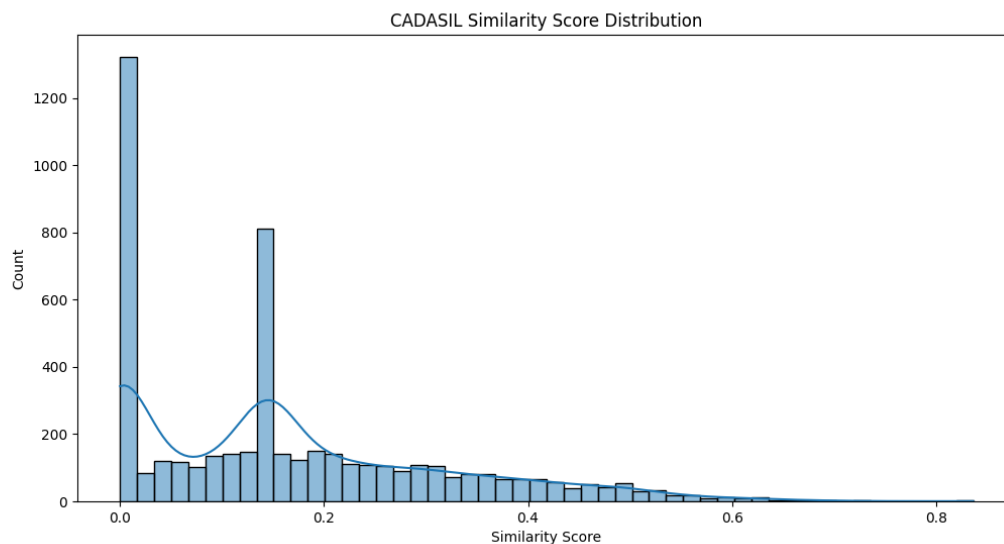
chromosomal and statistical details. The variants rs117415762, rs6941837, rs72704785, rs11138313, and rs10745793 have the highest similarity scores, exceeding 0.7 (70%).



**Figure 2: Top 100 predicted SNPs.** The red line represents a similarity threshold corresponding to a score of 70%. SNPs above this threshold are annotated

These SNPs are the result of precise filtering based on similarity scores, which eliminated a set of low-scoring SNPs and retained only those with

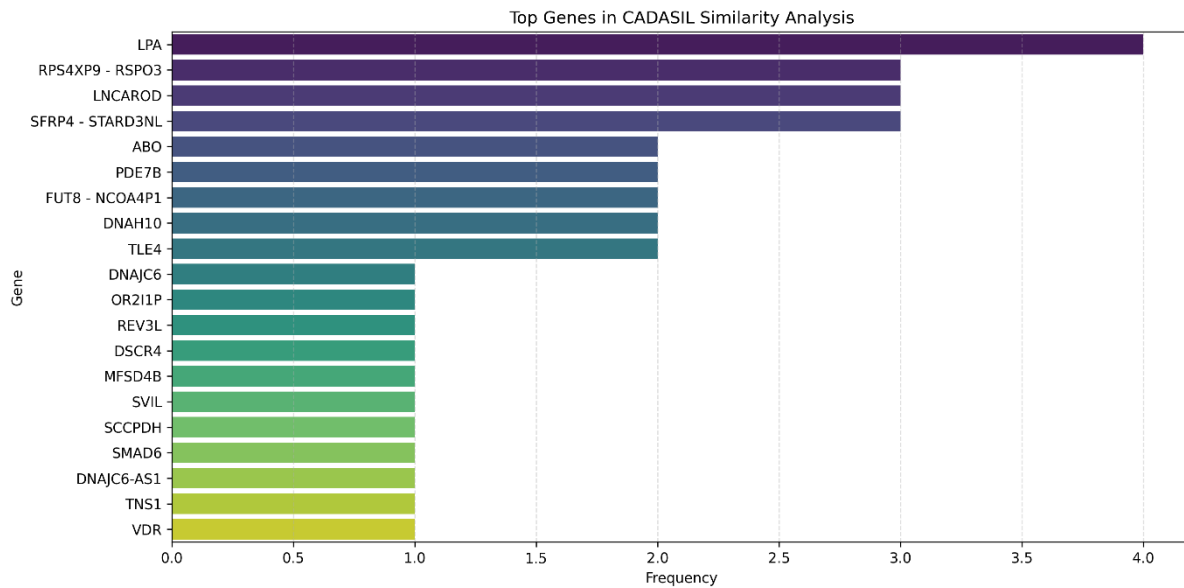
significant similarity scores. Figure 3 shows the variations in genomic similarity scores as a function of SNP.



**Figure 3: Distribution of genetic similarity scores for candidate SNPs for CADASIL**

The distribution shows that the majority of variants have low scores, while an extended distribution tail reveals a subset of SNPs with high genetic similarity, potentially involved in the pathogenesis of the disease. This subset presents the 100 candidate SNPs shown in Figure 2.

Finally, to understand the involvement of these SNPs in the mechanisms of CADASIL, we identified the most frequent genes (predicted at least once). These genes carry certain SNPs among the 100 predicted. Figure 4 shows these gene frequency results and Table 1 presents their functions extracted from NCBI gene [10].



**Figure 4. Frequency of genes carrying the most predictive SNPs for CADASIL**

**Table 1. Frequent Genes Carrying Predicted SNPs and Their Functional Annotations**

Gene	Frequency	Function
Lipoprotein(a) ( <i>LPA</i> )	4	Apo(a) is a serine protease component of lipoprotein(a) that inhibits tissue-type plasminogen activator, promotes thrombogenesis, and is linked to atherosclerosis.
ribosomal protein S4X pseudogene 9 ( <i>RPS4XP9</i> )	3	ribosomal protein S4X pseudogene 9
R-spondin 3 ( <i>RSPO3</i> )	3	This gene encodes a regulator of Wnt/ $\beta$ -catenin and PCP signaling, influencing development and cell growth. It is associated with bone mineral density, fracture risk, and may contribute to tumorigenesis.
lncRNA activating regulator of <i>DKK1</i> ( <i>LNCAROD</i> )	3	lncRNA activating regulator of <i>DKK1</i>
secreted frizzled related protein 4 ( <i>SFRP4</i> )	3	<i>SFRP4</i> is a secreted modulator of Wnt signalling; it contains a Frizzled-like Wnt-binding domain and its expression in ventricular myocardium correlates with apoptosis-related gene expression.
STARD3 N-terminal like ( <i>STARD3NL</i> )	3	This gene encodes a late-endosomal protein with a MENTAL domain, involved in cholesterol binding and endosomal cholesterol transport.
alpha 1-3-N-acetylgalactosaminyltransferase and alpha 1-3-galactosyltransferase ( <i>ABO</i> )	2	<i>ABO</i> encodes glycosyltransferases that determine blood group by converting the H antigen into A or B antigens; the O group results from a frameshift mutation.
phosphodiesterase 7B ( <i>PDE7B</i> )	2	<i>PDE7B</i> encodes a cAMP-specific phosphodiesterase that regulates signaling by hydrolyzing cAMP to 5'-AMP.
fucosyltransferase 8 ( <i>FUT8</i> )	2	This gene encodes an enzyme belonging to the family of fucosyltransferases. The product of this gene catalyzes the transfer of fucose from GDP-fucose to N-linked type complex glycopeptides. This enzyme is distinct from other fucosyltransferases which catalyze alpha1-2, alpha1-3, and alpha1-4 fucose addition. The expression of this gene may contribute to the malignancy of cancer cells and to their invasive and metastatic capabilities. Alternative splicing results in multiple transcript variants.
nuclear receptor coactivator 4 pseudogene 1 ( <i>NCOA4P1</i> )	2	

dynein axonemal heavy chain 10 ( <i>DNAH10</i> )	2	Dyneins are microtubule-associated motor protein complexes composed of several heavy, light, and intermediate chains. The axonemal dyneins, found in cilia and flagella, are components of the outer and inner dynein arms attached to the peripheral microtubule doublets. <i>DNAH10</i> is an inner arm dynein heavy chain
TLE family member 4, transcriptional corepressor ( <i>TLE4</i> )	2	TLE4 is a transcription corepressor involved in negative regulation of Wnt signaling and RNA polymerase II-mediated transcription.
DnaJ heat shock protein family (Hsp40) member C6 ( <i>DNAJC6</i> )	1	<i>DNAJC6</i> is a DNAJ/HSP40 family chaperone protein that stimulates ATPase activity and contains J, G/F-rich, and cysteine-rich domains.
olfactory receptor family 2 subfamily I member 1, pseudogene ( <i>OR211P</i> )	1	<i>OR211P</i> is an olfactory receptor, a GPCR that detects odorants and initiates neuronal signaling; part of the largest gene family in the genome.
REV3 like, DNA directed polymerase zeta catalytic subunit ( <i>REV3L</i> )	1	<i>REV3L</i> encodes the catalytic subunit of DNA polymerase zeta, involved in translesion DNA synthesis and mitochondrial DNA protection; mutations can cause Mobius syndrome
Down syndrome critical region 4 ( <i>DSCR4</i> )	1	<i>DSCR4</i> is located in a Down syndrome-associated region of chromosome 21 and is transcribed from a bidirectional retroviral promoter.
solute carrier family 60 member 2 ( <i>MFSD4B</i> )	1	<i>MFSD4B</i> is a predicted symporter involved in sodium ion and transmembrane transport, localized to the apical plasma membrane.
supervillin ( <i>SVIL</i> )	1	<i>SVIL</i> encodes a bipartite actin-binding protein linking the cytoskeleton to the plasma membrane, involved in myosin II assembly and focal adhesion dynamics.
saccharopine dehydrogenase ( <i>SCCPDH</i> )	1	<i>SCCPDH</i> is a predicted oxidoreductase involved in glycolipid biosynthesis, localized to lipid droplets and the midbody.
SMAD family member 6 ( <i>SMAD6</i> )	1	<i>SMAD6</i> is a signal-transducing transcriptional modulator that negatively regulates BMP and TGF- $\beta$ /activin signaling.
DnaJ heat shock protein family (Hsp40) member C6 ( <i>DNAJC6</i> )	1	<i>DNAJC6</i> is a DNAJ/HSP40 chaperone protein that stimulates ATPase activity and contains J, G/F-rich, and cysteine-rich domains.
prostaglandin D2 receptor ( <i>ASI</i> (Also known as <i>PTGDR</i> ))	1	<i>ASI</i> encodes a GPCR that acts as a receptor for prostaglandin D2, mediating allergic inflammation and asthma-related signaling.
tensin 1 ( <i>TNSI</i> )	1	<i>TNSI</i> encodes a focal adhesion protein that crosslinks actin filaments, contains an SH2 domain, and is a calpain II substrate.
vitamin D receptor ( <i>VDR</i> )	1	<i>VDR</i> encodes the vitamin D3 receptor, a nuclear hormone receptor regulating mineral metabolism, immune response, and cancer-related pathways; mutations can cause type II vitamin D-resistant rickets.

## 4. Discussion

### 4.1. Predicted SNP

Our study has identified 100 SNPs associated with CADASIL. The similarity score provides statistical validation of the predictions. The identified SNPs have a similarity score ranging from 0.534 to 0.836. According to Figure 2, rs117415762, rs6941837, rs72704785, rs11138313, and rs10745793 are the

most statistically significant SNPs with a similarity score exceeding 0.7. To date, no studies have established a direct link between these SNPs and CADASIL, making them new biomarkers for early diagnosis.

### 4.2. Predicted Genes

Up till now, most SNPs linked to CADASIL are located in the *NOTCH3* gene, which encodes a transmembrane protein primarily expressed in

vascular smooth muscle cells (VSMCs) and pericytes. This protein plays a key role in vascular development and function. Mutations in this gene can disrupt this function, which explains its direct link to this disease, characterized mainly by recurrent ischemic strokes [11].

Nevertheless, the pathological mechanisms involve multiple interdependent molecular interactions that control the expression of the genes in question and the cellular signaling of the translated proteins. CADASIL is not an exception; if its pathological mechanism were simply a disruption of *NOTCH3* function, prevention would be a simple matter of compensating for this disruption. However, this is not the case; to date, no preventive therapeutic treatments exist for individuals with these mutations [1].

In this context, our study proposed 24 genes (Figure 4) exhibiting variable prediction frequencies which may enhance our comprehension of the molecular mechanisms of this pathology.

Starting with the most frequent gene (n=4), which is *LPA*, this gene plays a role in lipid metabolism and is associated with the risk of cardiovascular disease. This link between *LPA* and CADASIL is confirmed by the work of Gong et al [12].

The second frequency level (n=3) identifies five genes: *RPS4XP9*, *LNCAROD*, *STARD3NL*, *RSPO3* and *SFRP4*. The prevalence of these genes makes them likely candidates for this pathology.

The molecular activities of these genes give us several hypotheses about their link with CADASIL. To begin with, we focus on *RPS4XP9* which is a pseudogene that may participate in the transcriptional and epigenetic regulation of genes involved in the disease pathway [13].

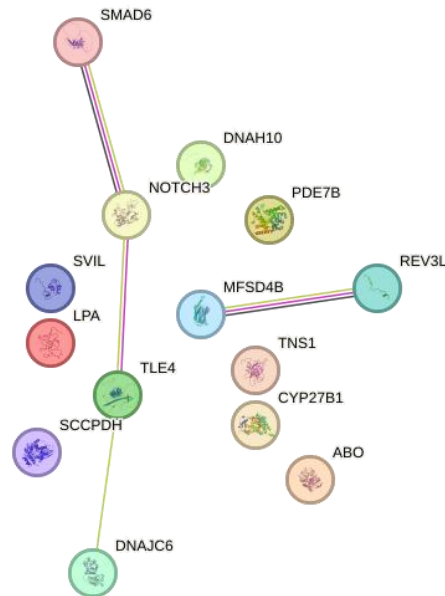
This regulation is also supported by *LNCAROD*, a long non-coding RNA (lncRNA) that may influence the expression of other genes involved in the disease pathway. The third gene in this group is *STARD3NL*. It contributes to intracellular cholesterol trafficking and the maintenance of membrane lipid equilibrium [10]. Therefore, any dysregulation of this gene can disrupt vascular membranes and promote CADASIL symptoms.

Thus, *RSPO3* and *SFRP4* also seem to be linked to CADASIL mechanisms, as they act as antagonistic regulators of the Wnt pathway to modulate vascular stability and angiogenesis [10].

At the third level of frequency, we find six genes: *ABO*, *PDE7B*, *FUT8*, *TLE4*, *NCOA4P1* and *DNAH10*. They are considered new candidates for this pathology, but with a lower probability than the first and second groups.

Finally, the fourth group, which is the least frequent, is composed of the following genes: *DNAJC6*, *OR21IP*, *REV3L*, *DSCR4*, *MFSD4B*, *SVIL*, *SCCPDH*, *SMAD6*, *DNAJC6-AS1*, *TNS1*, and *VDR*. These latter are also likely candidates for CADASIL with a lower probability than the previous groups. Carluccio et al have established the link between *VDR* and CADASIL, which supports our findings [14]. The other genes are exclusively associated with CADASIL in this study.

In relation with the association between the predicted genes and CADASIL, STRINGdb [15] confirmed that *NOTCH3*, the main gene for this pathology, directly interacts with *SMAD6*, *TLE4*, *DNAJC6* at a threshold of 0.15. It also indirectly interacts with all translated genes according to the same threshold and via a 50 enrichment of first and second level genes. This interaction is shown in Figure 5 below.



**Figure 5: Gene interaction networks proposed by STRINGdb between *NOTCH3* and predicted genes according to a threshold of 0.15 and without enrichment**

This compatibility between the documented genetics of CADASIL and the predicted genes confirms both the reliability of our results and the suitability of our methodological approach for studying rare diseases.

### 4.3. Limitations

Identifying the methodological limitations of any study is an important step. These limitations represent areas for improvement in future work. In our case, future improvements may concern the vector presentation model and prediction indicators. In fact, we chose a model that is compatible with the principle of genetic approximation and independent of costly infrastructure. However, using a model that is purely adapted to genetic contexts can improve our results. The prediction principle used requires genetic position indicators. But the use of other secondary functional indicators can increase the accuracy of predictions. These improvements require more expensive hardware resources and increase analysis time, but they can improve prediction. Finally, as with all *in silico* studies, the predictions generated require *in vitro* validation to confirm their biological role and establish their true significance for CADASIL.

## 5. Conclusion

Our FSL-based study enabled the prediction of 100 SNPs and 24 genes as biomarkers for CADASIL. These SNPs are excellent variants for early diagnosis. The interaction of the identified genes can build or complete cellular signaling pathways. These data can therefore provide an in-depth explanation of

the mechanisms of CADASIL and facilitate the development of therapeutic strategies. This study also confirms the compatibility of FSL in contexts characterized by a lack of data. This is a common and major hurdle in rare disease research. Consequently, our approach could be applied to other rare diseases with similar shortages of genetic information.

## 6. Acknowledgments

We thank Sarah Aguerd for her valuable assistance in improving the linguistic style and clarity of several paragraphs in the manuscript.

## 7. Authors Contributions

- **Ali Aguerd:** Conceptualization, Investigation, Methodology, Writing – Original Draft.
- **Oumaima Anachad:** Review & Editing, Validation.
- **Asmae Taheri:** Review & Editing, Validation.
- **Faïza Bennis:** Supervision, Review & Editing.
- **Fatima Chegdani:** Project Administration, Supervision, Validation, Review & Editing.

All authors read and approved the final manuscript.

## 8. Statements and Declarations

### Ethical considerations

This study did not involve human participants or animals; therefore, ethical approval was not required.

### Consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Declaration of conflicting interest

The authors declare that they have no competing interests.

### Funding statement

No funding was received for this work.

## 9. Availability of data

The results of this study are available as a zip file named “**Predicted-Data**” which contains 2 files: The 100 predicted SNPs and their details, and the Python script used.

## 10. References

- [1] Y. Yamamoto, Y.-C. Liao, Y.-C. Lee, M. Ihara, and J. C. Choi, “Update on the Epidemiology, Pathogenesis, and Biomarkers of Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy,” *J. Clin. Neurol.*, vol. 19, no. 1, pp. 12–27, Jan. 2023, doi: 10.3988/jcn.2023.19.1.12.
- [2] F. Felix-Ilemhembhio, K. Kocsy, M. Azzouz, and A. Majid, “The role of NOTCH3 in CADASIL pathogenesis: insights into novel therapies,” *Brain Res.*, vol. 1863, p. 149754, May 2025, doi: 10.1016/j.brainres.2025.149754.
- [3] C. Opherck *et al.*, “Genome-wide genotyping demonstrates a polygenic risk score associated with white matter hyperintensity volume in CADASIL,” *Stroke*, vol. 45, no. 4, pp. 968–972, Apr. 2014, doi: 10.1161/STROKEAHA.113.004461.
- [4] M. Modat, D. M. Cash, L. Dos Santos Canas, M. Bocchetta, and S. Ourselin, “Machine Learning for Alzheimer’s Disease and Related Dementias,” in *Machine Learning for Brain Disorders*, vol. 197, O. Colliot, Ed., in *Neuromethods*, vol. 197. , New York, NY: Springer US, 2023, pp. 807–846. doi: 10.1007/978-1-0716-3195-9\_25.
- [5] Y. Ge, Y. Guo, S. Das, M. A. Al-Garadi, and A. Sarker, “Few-shot learning for medical text: A review of advances, trends, and opportunities,” *J. Biomed. Inform.*, vol. 144, p. 104458, Aug. 2023, doi: 10.1016/j.jbi.2023.104458.
- [6] A. Abdellaoui, L. Yengo, K. J. H. Verweij, and P. M. Visscher, “15 years of GWAS discovery: Realizing the promise,” *Am. J. Hum. Genet.*, vol. 110, no. 2, pp. 179–194, Feb. 2023, doi: 10.1016/j.ajhg.2022.12.011.
- [7] H. Liang, J. C. Sedillo, S. J. Schrodi, and A. Ikeda, “Structural variants in linkage disequilibrium with GWAS-significant SNPs,” *Heliyon*, vol. 10, no. 11, June 2024, doi: 10.1016/j.heliyon.2024.e32053.
- [8] M. Cerezo *et al.*, “The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity,” *Nucleic Acids Res.*, vol. 53, no. D1, pp. D998–D1005, Jan. 2025, doi: 10.1093/nar/gkae1070.
- [9] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” 2019, *arXiv*. doi: 10.48550/ARXIV.1908.10084.
- [10] G. R. Brown *et al.*, “Gene: a gene-centered information resource at NCBI,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D36–D42, Jan. 2015, doi: 10.1093/nar/gku1055.
- [11] L. Yuan, X. Chen, J. Jankovic, and H. Deng, “CADASIL: A NOTCH3-associated cerebral small vessel disease,” *J. Adv. Res.*, vol. 66, pp. 223–235, Dec. 2024, doi: 10.1016/j.jare.2024.01.001.
- [12] M. Gong *et al.*, “Clinical and genetic features in a family with CADASIL and high lipoprotein (a) values,” *J. Neurol.*, vol. 257, no. 8, pp. 1240–1245, Aug. 2010, doi: 10.1007/s00415-010-5496-5.
- [13] S. H. Qian, L. Chen, Y.-L. Xiong, and Z.-X. Chen, “Evolution and function of

- developmentally dynamic pseudogenes in mammals,” *Genome Biol.*, vol. 23, no. 1, p. 235, Nov. 2022, doi: 10.1186/s13059-022-02802-y.
- [14] M. A. Carluccio *et al.*, “Vitamin D levels in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL),” *Neurol. Sci.*, vol. 38, no. 7, pp. 1333–1336, July 2017, doi: 10.1007/s10072-017-2900-2.
- [15] D. Szklarczyk *et al.*, “The STRING database in 2025: protein networks with directionality of regulation,” *Nucleic Acids Res.*, vol. 53, no. D1, pp. D730–D737, Jan. 2025, doi: 10.1093/nar/gkae1113.