Teaching Machines to Spot Lupus: How AI Is Cracking One of Medicine's Toughest Diagnostic Puzzles

When Your Immune System Becomes Your Enemy

Imagine your body's defense system suddenly turning against you, attacking your own organs as if they were foreign invaders, creating a dangerous form of “friendly fire”. This is the reality for millions of people living with Systemic Lupus Erythematosus (SLE), commonly known as lupus, a chronic autoimmune disease that can affect virtually any organ system in the body (Ameer *et al.* 2022).

The Diagnostic Challenge

Lupus is notoriously difficult to diagnose. It's often called "the great imitator" because its symptoms, ranging from joint pain and skin rashes to kidney problems and neurological issues, can mimic many other conditions.

The current diagnostic process relies heavily on detecting anti-nuclear antibodies (ANAs) in blood tests. However, here's the catch: while more than 95% of lupus patients test positive for ANA, so do many people with other autoimmune conditions or even healthy individuals (Ameer *et al.* 2022). This overlap creates a diagnostic puzzle that can delay proper treatment for years (Mitchell 2024).

Enter Machine Learning: A Digital Detective

Researchers from Taiwan tackled this challenge by training six different machine learning algorithms to distinguish lupus patients from non-lupus individuals, all of whom tested positively for ANA. Think of these algorithms as highly sophisticated pattern-recognition systems that can identify complex relations between many variables, which might be impossible to spot, even for experts.

The study analyzed data from 2,838 participants: 946 with confirmed lupus and 1,892 without the disease (Chung *et al.* 2024). These algorithms had been trained on a wide range of information:

- Clinical data from electronic health records
- 684 single-nucleotide polymorphisms (SNPs) - These are tiny variations in DNA sequences that can influence disease susceptibility.
- Polygenic risk scores - estimates person's genetic predisposition to lupus.

Among the six machine learning models tested, two turned out to be the most accurate. Namely, the Random Forest and Extreme Gradient Boosting (XGB). Both models achieved impressive accuracy rates, with the XGB model reaching an evaluation score of 0.87, indicating strong discriminatory ability (Chung *et al.* 2024).

Which features point to lupus?

The next step was to identify which variables were most important for the models' prediction. This required the usage of a technique called SHAP (Shapley Additive Explanations), which points out the most influential variables. The leading diagnostic features included:

- Anti-double strand DNA antibodies: Specific autoantibodies highly associated with lupus
- ANA titer levels: Indicate the concentration of antinuclear antibodies in the blood
- ANA patterns: Specific visual patterns seen under microscopy, particularly the AC4 pattern
- Complement levels (C3 and C4): Proteins involved in immune responses that are often depleted in active lupus
- Specific SNPs
- Polygenic risk scores

Potential

One of the key strengths of this work lies in its clinical potential. Doctors often struggle to know which patients to monitor more closely (Mitchell 2024). This kind of model could function in a rheumatologist's workflow as an intelligent assistant: flagging high-risk patients early, guiding further testing, and reducing the burden of unnecessary exams for low-risk individuals.

In practice, this means faster diagnosis, fewer false alarms, and more efficient use of medical resources. Over time, such tools could become part of personalized, precision medicine strategies, supporting doctors with data-driven risk stratification rather than relying on partial clinical intuition.

Novel Discoveries

Perhaps the most intriguing finding is the identification of eleven SNPs associated with SLE in a subgroup of patients with very high ANA titers (≥ 1:640) (Chung *et al.* 2024). Six of these SNPs showed a positive association (meaning they were more common in lupus patients), while five showed a negative association (less common in lupus patients). What makes this especially novel is that many of these genetic variants come from genes not previously linked with SLE. For example:

- VIT, involved in iron transport and antioxidant activity (Aghabi *et al.* 2021)
- SLC7A2, a transporter related to inflammatory responses (Cloots *et al.* 2017)
- AUTS2, more commonly studied in neurodevelopment (Hori *et al.* 2021)
- PACRG, linked to Parkinson's disease and infection susceptibility (Taylor *et al.* 2009)

These new SNPs act like genetic clues switches in the genome that might tilt someone's immune system toward or away from developing lupus. Their discovery opens up fresh lines of investigation: Why do these genes matter in autoimmunity? Could they inform new drug targets? Can they improve risk scores?

Limitations & What's Needed Next

Despite the excitement, there are important caveats. First, the study is retrospective, meaning it looks backward at existing data. Real-world clinical data in EHRs can be messy, incomplete, or biased, which can affect the model's reliability. Second, the researchers did not include other potentially rich data types like transcriptomics, proteomics, or cytokine profiles which might offer deeper biological insight. Third, and perhaps most critically, the entire study population is Taiwanese, meaning the results may not generalize well to other ethnic groups or geographic populations.

Moreover, while ML can highlight associations, it does not prove causality. The newly identified SNPs are strong candidates for further research, but we need functional lab studies to understand how (or if) they actually influence lupus risk.

To move forward, future work should focus on validating the model in diverse populations, running prospective studies where predictions are tested in real time, and doing biological experiments to test how the implicated genes work in the immune system. Ultimately, these steps are essential before such predictive models can be safely and effectively integrated into everyday clinical practice (Chung *et al.* 2024).

Machine learning diagnostics in society

Being able to diagnose patients more efficiently would increase throughput in the healthcare system. This would help lighten the burden at hospitals/clinics where lack of sufficient staff and being overcrowded are constant issues (Stale 2024).

While running a trained model is itself not very expensive, the cost of tests for producing the data may prevent this method from benefitting low-income countries and people of low socioeconomic status. If the future of diagnostics is more expensive, this would widen the gap in health and wealth across society. Further raising the threshold of healthcare will harm already vulnerable minority groups. (Hoagland & Kipping 2024) This increased cost factor could potentially be offset by a reduced total cost due to narrowing down the list of tests to the most beneficial ones.

Reanalyzing existing data, as was done in this study, is a sustainable research practice and aligns with the principles of Findability, Accessibility, Interoperability, and Reuse of digital assets (Wilkinson *et al.* 2016).

Ethical considerations

In this study, some genes identified as indicators of being positive/negative for SLE were also associated with other diseases. This expansion of analyzing a patient's genome is not necessarily something that all patients or their relatives want to know about. Informed consent should then include if the patient wants to be informed about a predisposition for certain illnesses. Consent is a continuous arrangement and must be possible to withdraw or modify after the initial signing of forms.

As a part of the analysis, a polygenic risk score was calculated. This statistical estimate indicates the risk of developing a certain disease. When this information is provided to insurance companies it has resulted in increased premiums and refusal of service etc. (Yanes *et al.* 2024) This type of discrimination of a still healthy person based on genetics is likely to become more prevalent the more these risk scores are used in diagnostics.

In the wake of the rapid development of various AI tools, institutions have been slow to produce updated guidelines and regulations. This gap is slowly being filled as new legal documents take effect and countries and other organizations implement them. (European Union 2024, Vardas *et al.* 2025)

Contributions

Jan Indrzejczak:

- When Your Immune System Becomes Your Enemy
- The Diagnostic Challenge
- Enter Machine Learning: A Digital Detective
- Which features point to lupus?

Md Mehedi Hassan:

- Potential
- Novel Discoveries
- Limitations & What's Needed Next

Rasmus Hammar:

- Machine learning diagnostics in society
- Ethical considerations

All authors reviewed and approved the final text.

References

Aghabi D, Sloan M, Dou Z, Guerra AJ, Harding CR. 2021. The vacuolar iron transporter mediates iron detoxification in Toxoplasma gondii. 2021.09.08.458725.

Ameer MA, Chaudhry H, Mushtaq J, Khan OS, Babar M, Hashim T, Zeb S, Tariq MA, Patlolla SR, Ali J, Hashim SN, Hashim S, Ameer MA, Chaudhry H, Mushtaq J, Khan OS, Babar M, Hashim T, Zeb S, Tariq MA, Patlolla SR, Ali J, Hashim SN, Hashim S. 2022. An Overview of Systemic Lupus Erythematosus (SLE) Pathogenesis, Classification, and Management. Cureus, doi 10.7759/cureus.30330.

Chung C-W, Chou S-C, Hsiao T-H, Zhang GJ, Chung Y-F, Chen Y-M. 2024. Machine learning approaches to identify systemic lupus erythematosus in anti-nuclear antibody-positive patients using genomic data and electronic health records. BioData Mining 17: 1.

Cloots RHE, Sankaranarayanan S, Poynter ME, Terwindt E, van Dijk P, Lamers WH, Eleonore Köhler S. 2017. Arginase 1 deletion in myeloid cells affects the inflammatory response in allergic asthma, but not lung mechanics, in female mice. BMC Pulmonary Medicine 17: 158.

European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).

Hoagland A, Kipping S. 2024. Challenges in Promoting Health Equity and Reducing Disparities in Access Across New and Established Technologies. Canadian Journal of Cardiology 40: 1154–1167.

Hori K, Shimaoka K, Hoshino M. 2021. AUTS2 Gene: Keys to Understanding the Pathogenesis of Neurodevelopmental Disorders. Cells 11: 11.

Mitchell JL. 2024. Understanding the impact of delayed diagnosis and misdiagnosis of systemic lupus erythematosus (SLE). Journal of Family Medicine and Primary Care 13: 4819.

Stale SR. 2024. Läkarförbundet granskar: Vårdplatsbristen. The Swedish Medical Association

Taylor JM, Wu R-M, Farrer MJ, Delatycki MB, Lockhart PJ. 2009. Analysis of PArkin Co-Regulated Gene in a Taiwanese-ethnic Chinese cohort with early-onset Parkinson's disease. Parkinsonism & Related Disorders 15: 417–421.

Vardas EP, Marketou M, Vardas PE. 2025. Medicine, healthcare and the AI act: gaps, challenges and future implications. European Heart Journal Digital Health 6: 833–839.

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL,

Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018.

Yanes T, Tiller J, Haining CM, Wallingford C, Otlowski M, Keogh L, McInerney-Leo A, Lacaze P. 2024. Future implications of polygenic risk scores for life insurance underwriting. npj Genomic Medicine 9: 25.