

# Soren Basnet

412 Wisconsin Ave, Madison, WI | basnetsoren01@gmail.com | 859-436-3520 | linkedin.com/in/sorenbasnet

## Education

---

**Luther College** | Decorah, IA

Graduation : May 2025

**B.A. in Computer Science, Minor in Mathematics**

**Honors:** Dean's List (2 Semesters), Dean's Scholarship

**Relevant Courses:** Advanced Algorithms and Data Structures, Data Modeling and Querying, Object-Oriented Programming With Java, Applied Machine Learning, Probability Theory, Experimental Design and ANOVA

**University of Wisconsin-Madison** | Madison, WI

Expected Graduation: May 2027

**M.S. Statistics and Data Science**

**Relevant Courses:** Statistical Methods I, Introduction to Statistical Inference, Classification and Regression Trees, Statistical Learning

## Research and Professional Experience

---

**Data Scientist / Biostatistician Intern**, Mayo Clinic | Rochester, MN

**Jun 2026 – Present**

- Draft scientific proposal and present study design and findings to clinicians and statisticians, incorporating peer feedback emphasizing collaboration.
- Design reproducible statistical analyses and predictive models including logistic regression and Cox proportional hazards model to investigate data-driven research on genetic contribution to infection severity.
- Process large, structured dataset pipelines using R, Python and Google Cloud Processing for statistical analyses.
- Construct risk scores to study genetic contribution to infections by integrating published genetic weights and evidence from existing odds ratios from association studies.

**Biostatistician Intern**, Mayo Clinic | Rochester, MN

**Feb 2024 – Aug 2025**

- Developed exploratory data analysis and preprocessing pipeline in R that was adopted by collaborators to predict recurring infection using ensemble machine learning (XGBoost, Random Forest, KNN and others) in Python.
- Conducted Genome-Wide Association Studies (GWAS) using PLINK, REGENIE and Bash on distributed computing environments ( Slurm/HTCondor ) to identify statistically significant genetic variants associated with disease risk using 10000+ genomic samples.
- Estimated genetic heritability rate using statistical tool : GCTA, validating a consistent rate of 32% for CLL patients.
- Co-authored a peer-reviewed publication in *Blood* (2024) on large-scale genomic analysis of over 1000 participants.

## Projects

---

**Predictive Analytics and Survival Modeling**, Independent Project

**Jan 2026 – May 2026**

- Designed data pipelines using R to integrate 14 heterogeneous clinical, laboratory, and genomic datasets for EDA and predictive modeling.
- Utilized machine learning and statistical models including decision trees, random forests, GUIDE and survival analysis to identify and replicate key predictors of 5-year mortality of patients from prevalent literature.
- Built reproducible preprocessing workflows including feature engineering, automated data validation/transformation of ~20,000 genomic features for scalable model development.

**VCF Parser**, Independent Project

**Jul 2026 – Present**

- Developing an automated tool to parse and validate Variant Call Format (VCF) files to expedite ETL pipelines and fast variant searching.
- Standardizing genetic variant quality control metrics with automated tracking and change-logging for workflow reproducibility.
- Building a centralized platform for script and dataset documentation to track data provenance and compare output similarities across studies.

## Technical Skills

---

**Programming Languages:** Python ( Pandas,NumPy,Scikit-learn,PyTorch/TensorFlow ), R ( dplyr,ggplot2,arrow,rpart ), Bash,SAS

**Tools & Infrastructure:** Git, Linux/Unix, HTCondor/Slurm, Google Cloud Platform (GCP), AWS