

TOOBA ZAHRA

Bioinformatics Software Engineer | Research Software Engineering | Scientific Tool Building

Islamabad, Punjab, PK | tooba.zahra19@gmail.com | [GitHub](#) | [LinkedIn](#)

PROFESSIONAL SUMMARY

Bioinformatics graduate student (MS, expected Feb 2027) with strong expertise in Artificial Intelligence and Data Science, and hands-on experience in vaccine design, alongside applying AI/ML techniques to biological and healthcare datasets, from raw sequencing files to deployed, user-facing tools. Background spans machine learning for biomedical classification (Random Forest, scikit-learn), in-silico vaccine design, RNA-seq and differential expression analysis, and building reproducible pipelines (Nextflow, Docker) and interactive data tools (Streamlit, Plotly). Comfortable working across the full data science stack: preprocessing, analysis, model evaluation, and deployment. Currently combining this technical work with a Bioinformatics Sales Coordinator role at a Tech-Bio company.

TECHNICAL SKILLS

- Artificial Intelligence & Machine Learning: Classification algorithms (Random Forest), feature engineering, model evaluation and validation, applied to biological and biomedical datasets
- In-Silico Vaccine Design: Computational antigen/target analysis and vaccine design workflows, developed through MGBio-sponsored training
- Programming & Analysis: Python (Pandas, NumPy, scikit-learn, Matplotlib, Streamlit, Flask), R (Bioconductor, DESeq2, edgeR, ggplot, dplyr), C++
- Software Engineering: Building and deploying interactive web apps, custom file parsers, dependency/environment debugging, containerization
- Bioinformatics: RNA-seq analysis, differential gene expression, sequence alignment, metagenomics, sequencing QC, variant analysis basics, functional enrichment
- Pipelines & Tools: Nextflow, Docker, FastQC, MultiQC, STAR, HISAT2, Git/GitHub, Linux/Unix command line
- Deployment: Streamlit Community Cloud, GitHub Codespaces, cloud vs. local environment reconciliation
- Data Visualization & Statistics: Plotly, statistical analysis (ANOVA, Chi-square, LSD), model evaluation metrics

PROFESSIONAL EXPERIENCE

Bioinformatics Sales Coordinator | *MGBio Private Limited (Tech-Bio, NSTP, NUST)* | 2026 - Present

- Coordinate internship program operations including applicant data management and allocation across technical tracks for large applicant pools
- Support data compilation and outreach processes, working across spreadsheet and database tools to manage structured applicant and partner data
- Developed and manage the company website on WordPress
- Completed MGBio-sponsored "AI for Biologists 2.0" training on applying AI to biology; supported facilitation for other participants
- Completed MGBio-sponsored in-silico vaccine design training program (both programs through employee development scholarship)

Teaching Assistant | *COMSATS University Islamabad* | 2023 - 2024

- Assisted in teaching Bioinformatics Scripting to 50+ undergraduate students, providing hands-on coding support
- Conducted weekly lab sessions and graded 200+ assignments with detailed feedback, improving student programming outcomes by an average of 25%

KEY PROJECTS

Sequencing QC Dashboard (Streamlit) | [GitHub](#)

- Built and deployed an interactive web app that parses raw FASTQ files with a custom hand-written parser and generates quality metrics (read length, GC content, Phred quality scores)
- Designed automated PASS/WARNING/FAIL QC flagging with plain-language summaries, aimed at wet-lab scientists without command-line experience
- Resolved cross-environment dependency conflicts (Biopython, pyarrow, NumPy) between local development and cloud deployment

Breast Cancer Subtype Classifier (ML) | [GitHub](#)

- Applied AI/ML techniques (Random Forest, scikit-learn) to a healthcare classification problem, breast cancer subtype prediction, achieving 96.5% accuracy
- Handled data preprocessing, feature engineering, model evaluation, and validation on biological/biomedical datasets

Metagenomics Analysis Pipeline (Nextflow + Docker) | [GitHub](#)

- Built a reproducible, containerized metagenomics pipeline using Nextflow and Docker for multi-sample data processing
- Isolated each analysis tool in its own Docker container to eliminate dependency conflicts across environments

RNA-seq Nextflow + Docker Pipeline | [GitHub](#)

- Built an end-to-end RNA-seq pipeline (FastQC, Trimmomatic, STAR, featureCounts, MultiQC) with full Docker containerization
- Designed for reproducibility across compute environments, applicable to multi-parametric biomedical datasets

RNA-Seq Differential Expression Analysis: *C. elegans* Sex-Specific Genes | [GitHub](#)

- Performed end-to-end RNA-seq analysis of 6 samples achieving 76-84% alignment rates using STAR/HISAT2
- Identified sex-specific differentially expressed genes ($\text{padj} < 0.05$, $|\log_2\text{FC}| > 2$) from 19,985 genes using DESeq2 in R/Bioconductor

EDUCATION

MS in Bioinformatics | COMSATS University Islamabad, Pakistan | Expected Feb 2027

- CGPA: 3.7/4.0 (in progress)

BS in Bioinformatics | COMSATS University Islamabad, Pakistan | 2024

- CGPA: 3.55/4.0 | Class Rank: 2nd of 35 students
- Scotland Scholarship for Young Women & Girls 2022; Prime Minister's Youth Laptop Scheme (PMYLS) Recipient

LEADERSHIP

- Head of Graphic Design | IEEE SAC Team, COMSATS | 2022-2023 | Led design team for 15+ technical events
- Co-Head of Graphic Design | IDEA Society, COMSATS | 2022-2023 | Increased event engagement by 40%

REFERENCES

Dr. Farhan Haq | Tenured Associate Professor, Bio Science, CUI Islamabad Campus | farhan.haq@comsats.edu.pk

Mr. Muhammad Ammar | Managing Director, MGBio | ammar@mgbio.tech