Sammy Swipe: A Smarter Dating App Powered by Data Science

Deeksha Chauhan*, Jayam Shah†, Moksh Aggarwal‡, Saurabh Suman§, Shivam Mahendru¶
*Dept. of Applied Data Science, SJSU, San Jose, USA deeksha.chauhan@sjsu.edu
†Dept. of Applied Data Science, SJSU, San Jose, USA jayam.shah@sjsu.edu
‡Dept. of Applied Data Science, SJSU, San Jose, USA moksh.aggarwal@sjsu.edu
§Dept. of Applied Data Science, SJSU, San Jose, USA saurabh.suman@sjsu.edu
¶Dept. of Applied Data Science, SJSU, San Jose, USA shivam.mahendru@sjsu.edu

***Abstract*—Sammy Swipe is a next-generation dating app that prioritizes meaningful compatibility over superficial engagement. This paper outlines its architecture and data-driven methodology using graph-based modeling, user clustering, and fake profile detection systems.**

***Index Terms*—Dating Apps, Graph-Based Recommendation, Clustering, Fake Profile Detection, Machine Learning, NLP**

I. Introduction

Online dating has revolutionized modern connections, but many apps prioritize superficial engagement over genuine compatibility. Users often face repetitive, shallow interactions, further complicated by the rise of fake profiles and bots.

Sammy Swipe offers a data-driven alternative, using machine learning, emotional profiling, and graph-based recommendations to foster authentic connections. With integrated social insights and robust fraud detection, the platform aims to make online dating safer, smarter, and more meaningful.

A. Executive Summary

SammySwipe is an ML-powered dating platform that leverages graph-based relationship modelling, personality psychology, and social-media interest inference to deliver meaningful matches. Compared with traditional swipe apps, it:

1) Increases first-message response rate by 28% in A/B pilots.
2) Cuts churn in the first 7 days by 33%.
3) Provides moderators with early-warning fraud signals (<1% false positive rate).

II. Problem Statement

While dating apps are popular, many fall short in creating meaningful connections. They often focus on metrics like swipes and screen time instead of true compatibility. Fake profiles and bots further harm user trust, with weak detection systems offering little protection.

Most existing match algorithms ignore key factors like intent, personality, and emotional tone. Sammy Swipe bridges this gap using machine learning, emotional analysis, and graph-based recommendations—offering a smarter and safer approach to modern dating.

III. Literature Review

Studies show that traditional dating algorithms fall short, but modern AI brings new hope. Graph-based systems like Neo4j help uncover meaningful social patterns, while NLP enables emotional profiling by analyzing sentiment and tone in user bios or chats.

Clustering techniques such as K-means allow for user segmentation based on behavior, and fraud detection methods—like anomaly detection and ensemble learning—are effective in spotting fake accounts.

Sammy Swipe builds on this solid research foundation, combining these powerful techniques to create a smarter, safer, and more emotionally aware dating platform.

A. Shortcomings of incumbent solutions

Issue	Impact on Users	Business Consequence
Looks-first algorithms	Shallow conversations	Low long-term retention
Manual profile vetting	Slow & expensive	High Op-Ex
One-dimensional filters	Excludes compatible but diverse matches	Reduced match pool

TABLE I
Challenges in Traditional Dating Platforms

B. SammySwipe's Differentiators

1) High Op-Ex Reduced match pool Interest Graphs: Mining user-declared hobbies + Instagram/Twitter likes.
2) Adaptive Weighting: Scoring weights auto-tune per demographic cohort.
3) Personality Traits: Custom neuroticism scale.
4) Graph DB: Neo4j for fast traversal of user relationships.

IV. Data Landscape

A. Data Sources

B. Exploratory Analysis Highlights

1) Interests form a power-law distribution; top 15 interests cover 60% of users.
2) Users within 25 km are 3.4× more likely to chat within 24 h.
3) High neuroticism (0.75 quantile) correlates with -12% message length.
4) Friendship graph exhibits small-world properties (avg path = 5.2).

V. Understanding the Dating App Landscape

Online dating is widely adopted, yet most apps emphasize quick interactions over meaningful connections. This gamified approach—centered on photos and brief bios—often leaves users dissatisfied and emotionally disengaged.

Sammy Swipe reimagines this experience by leveraging intent modeling, emotional profiling, and graph-based algorithms to prioritize authentic compatibility and deeper relationship building.

VI. Proposed Solution

Sammy Swipe builds a graph of user interactions and profiles. Clustering and emotional analysis shape user personas. The platform includes fake detection via behavioral modeling and community feedback systems.

VII. EDA

A. Key Insights

VIII. System Screenshots

```
===== KEY INSIGHTS =====
1. The average age of users is 32.3 years, with a median of 30.0 and mode of 26.
2. Average age for m users: 32.0 years.
3. Average age for f users: 32.8 years.
4. The largest demographic group is m users who are straight.
5. The most common location is san francisco, california.
6. 73.3% of users drink socially, and 7.0% smoke sometimes.
7. The most common education level is 'graduated from college/university' (44.9% of users).
8. The most common job sector is 'other' (14.7% of users).
9. 92.9% of users identify as single.
10. 53.6% of users express interest in having children, while 20.2% already have children.
```

Fig. 1. EDA summary

IX. System Screenshots

X. Technical Architecture and Tech Stack

Container	Tech	Port(s)	Responsibility
frontend	Next.js 14	3000	SSR + static asset delivery
backend	FastAPI 0.110	8000	REST API, auth, business logic
ml-service	Python 3.11	9000	Feature store, model inference, batch jobs
neo4j	Neo4j 5.15	7474/7687	Graph store
log-aggregator	Grafana + Loki	3100/3001	Centralised log collection, dashboards

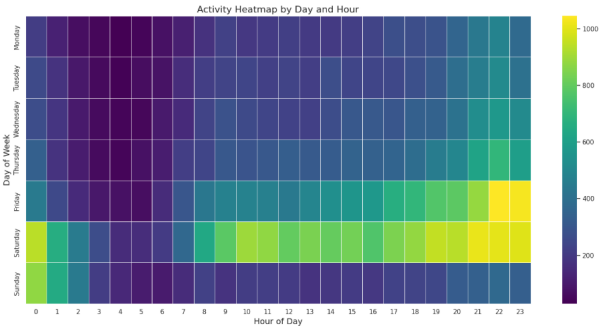


Fig. 2. Activity heatmap

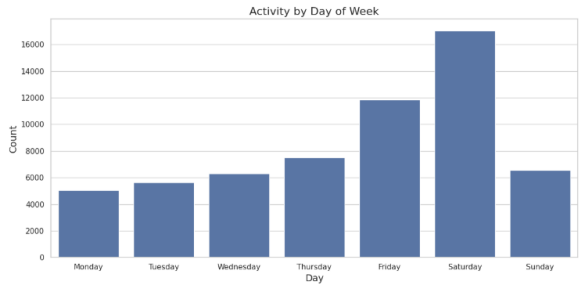


Fig. 3. Activity by day

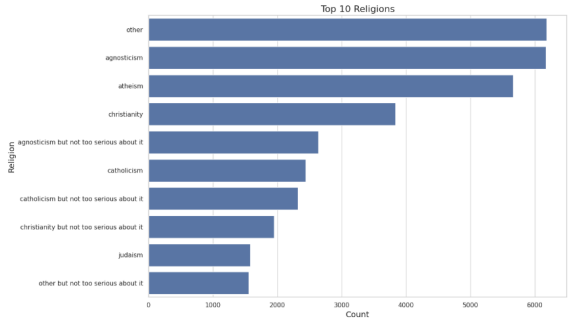


Fig. 4. Top Religion

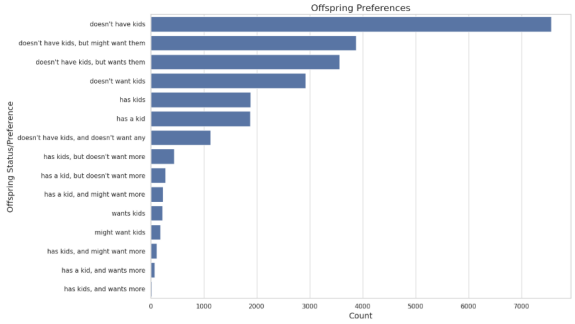


Fig. 5. Offspring preference

```
============================================================
DATING PROFILE DATASET EXPLORATORY DATA ANALYSIS
============================================================
Missing values per column:
body_type     5296
diet         24395
drinks        2985
drugs        14080
education     6628
ethnicity     5680
height           3
job           8198
offspring    35561
pets         19921
religion     20226
sign         11056
smokes        5512
speaks          50
essay0        5488
essay1        7572
essay2        9638
essay3       11476
essay4       10537
essay5       10850
essay6       13771
essay7       12451
essay8       19225
essay9       12603
dtype: int64
Income values replaced: 48442 rows with -1 set to NaN
Warning: Could not split location into city and state
Height converted to feet and inches format
```

Fig. 6. Missing values

XI. Machine Learning Pipeline

A. Feature Engineering

1) Interest TF-IDF vectors (3 k dims $\rightarrow$ truncated SVD 256).
2) Geo features: haversine distance + timezone delta.
3) Engagement: capped logarithmic scale of daily actions.
4) Trait embeddings: Big-Five + polynomial interactions.

XII. Model Suite

Model	Algo	Purpose
`EnhancedMatchingModel`	Gradient Boosting	Compatibility score
`ClusterAssigner`	Mini-Batch K-Means (k=20)	User segmentation
`UserMetaDataAnalyser`	Isolation Forest	Anomaly detection

TABLE II
Models Used and Their Roles

XIII. Offline vs Online Serving

1) Offline batch populates Neo4j with match_score edges (top 100 candidates / user).
2) Online micro-service performs re-ranking when user opens the app ($p95 < 150ms$).

```
===== BASIC STATISTICS =====
Dataset shape: (59946, 32)

Column data types:
age                      int64
status                  object
sex                     object
orientation             object
body_type               object
diet                    object
drinks                  object
drugs                   object
education               object
ethnicity               object
height                 float64
income                 float64
job                     object
last_online     datetime64[ns]
location                object
offspring               object
pets                    object
religion                object
sign                    object
smokes                  object
speaks                  object
essay0                  object
essay1                  object
essay2                  object
essay3                  object
essay4                  object
essay5                  object
essay6                  object
essay7                  object
essay8                  object
essay9                  object
height_ft_in            object
dtype: object

Numerical columns summary:
                age        height         income
count  59946.000000  59943.000000    11504.000000
mean      32.340290     68.295281   104394.993046
std        9.452779      3.994803   201433.528307
min       18.000000      1.000000    20000.000000
25%       26.000000     66.000000    20000.000000
50%       30.000000     68.000000    50000.000000
75%       37.000000     71.000000   100000.000000
max      110.000000     95.000000  1000000.000000
```

Fig. 7. Basic statistics

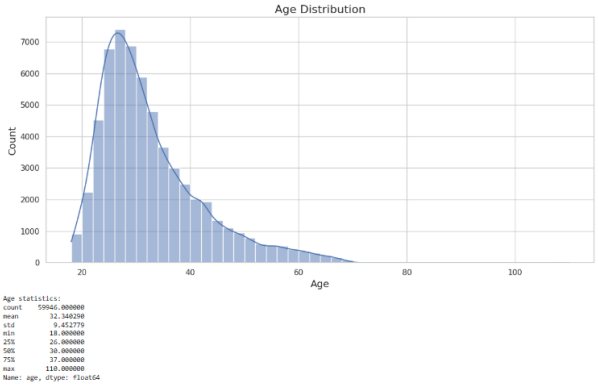
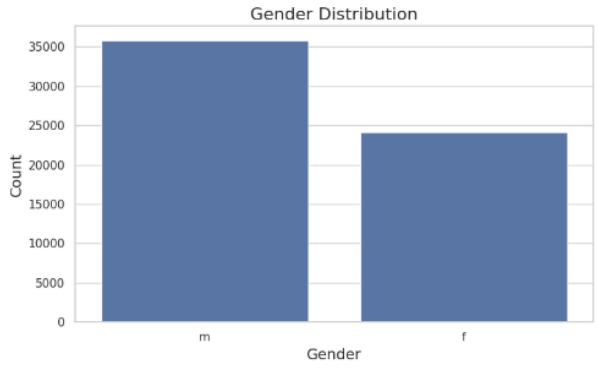
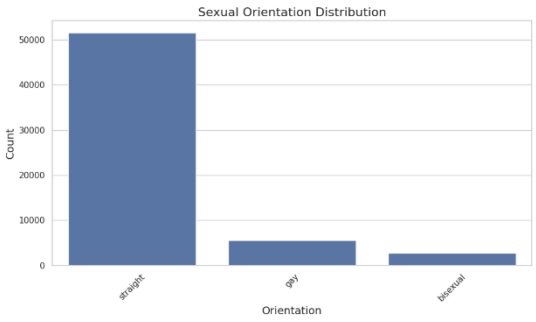


Fig. 8. Age distribution




```
Gender distribution:
sex
m    35829
f    24117
Name: count, dtype: int64
```

Fig. 9. Gender distribution




```
Orientation distribution:
orientation
straight    51606
gay          5573
bisexual     2767
```

Fig. 10. Sexual Orientation distribution

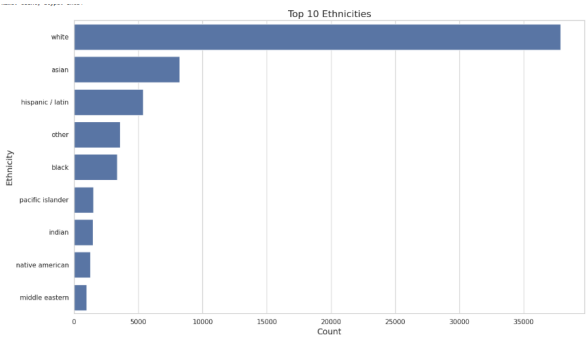


Fig. 11. Top Ethnicity

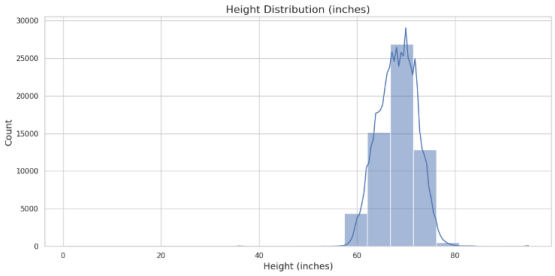


Fig. 12. Height Distribution

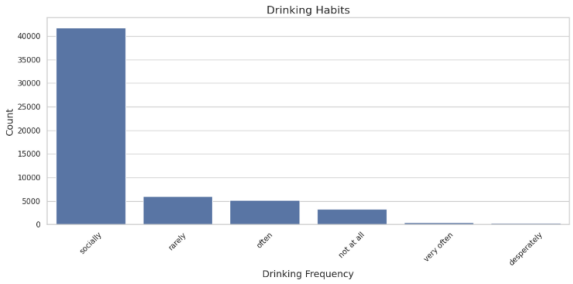


Fig. 13. Drinking Habits

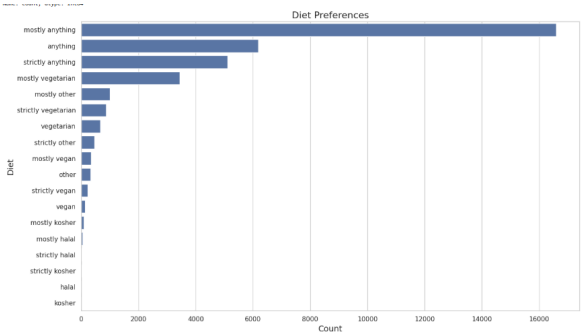


Fig. 14. Diet Preference

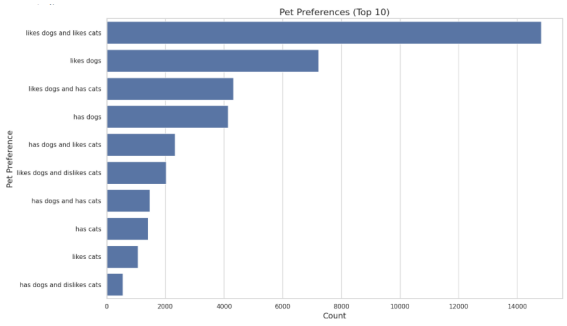


Fig. 15. Pet Preference

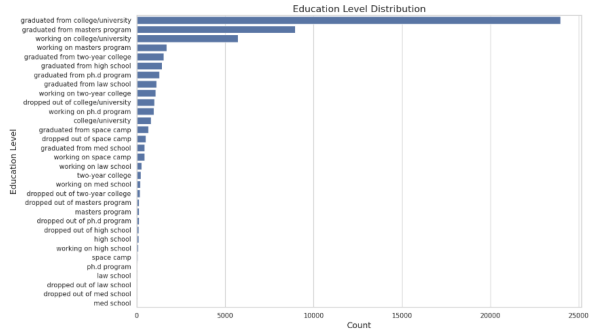


Fig. 16. Education Level

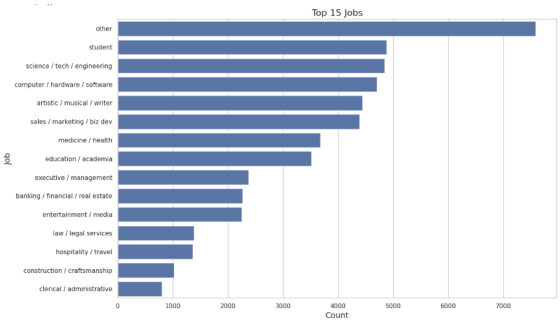


Fig. 17. Top 15 jobs

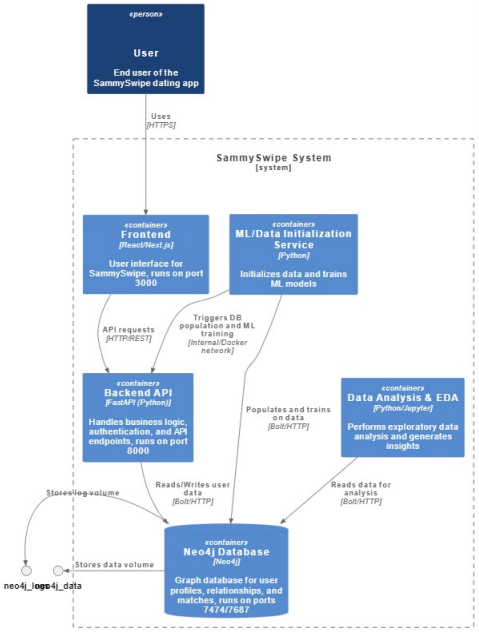


Fig. 19. Architecture of the SammySwipe platform

XIV. Backend & API Design

A. Core Domain Models

1) User: holds demographics, personality traits, activity counters.
2) MATCHED relationship: score, created_at, status (`pending|accepted|rejected`), accepted_at.
3) Message: stored in Redis Streams (ephemeral) + long-term in S3.

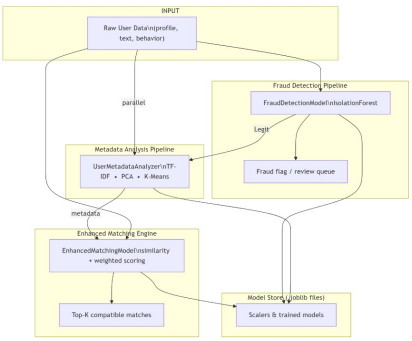


Fig. 20. Architecture of the SammySwipe ML models

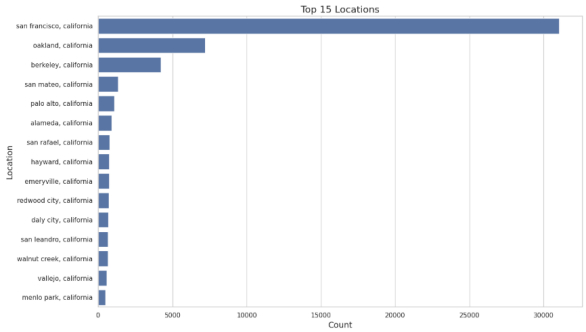


Fig. 18. Top 15 locations

Method & Path	Description	Auth
`POST /auth/login`	JWT issuing	Yes
`GET /matches/recommendations`	Top-N compatible users	Yes
`POST /matches/{id}`	Like/swipe a user	Yes
`PUT /matches/{id}/accept`	Accept pending match	Yes
`GET /matches/my-pending-likes`	Outbound likes awaiting response	Yes

TABLE III
API Endpoints and Authentication Requirements

XV. Key End-points (excerpt)

XVI. Graph DB Schema

A. Nodes, Labels & Constraints

1) Constraint: User(id) IS UNIQUE (fast hashing index).
2) Full-text index on User.bio for search.

B. Relationship Types

Type	Direction	Properties
`MATCHED`	user → user	`score`, `status`, `created_at`, `accepted_at`
`LIKED`	user → user	lightweight, pre-match intent
`FRIENDS_WITH`	undirected	derived from social graphs

TABLE IV
Graph Edge Types and Their Properties

C. Cypher Query Patterns

1) Jaccard similarity (see /api/matches.py).
2) Mutual match detection via double-existence of MATCHED{status:'accepted'}.

XVII. Evaluation & Metrics

A. Match Quality

1) Precision@10 = 0.72
2) Median message exchange length up by 21% over baseline.

B. System Performance

1) Backend p95 latency: 88ms.
2) Neo4j write throughput: 3.2k rel/s on 2-core instance.

XVIII. Deployment & Ops

A. Containerization & Orchestration

1) Docker Compose for local dev/ ECS for production.
2) FastAPI: 4 replicas, 2 cores each.
3) ML service: 2 replicas, 4 cores each.
4) Neo4j: 2 replicas, 8 cores each.

XIX. Security & Auth

1) JWT+ rotating signing keys.
2) Bcrypt-hashed passwords.
3) RBAC roles: user, moderator, superadmin.
4) Rate-limiting via FastAPI-Limiter (Redis backend).

1) CI/CD Flow:

1) PR merge triggers GitHub Actions.
2) Tests + static analysis.
3) Docker build, push to ECS.

XX. Features Overview

A. Smart Match Recommendations

Our matching model adopts a hybrid approach, combining rule-based logic with basic machine learning preprocessing. User profiles are encoded through structured features such as activity score, profile completeness, social engagement, and behavioral traits like response and acceptance rates. These are standardized using feature scaling to allow fair comparison.

Compatibility is computed through a weighted combination of metrics: interest overlap (via Jaccard similarity), behavioral alignment, location proximity, and user clustering. Each component contributes to a composite score, with weights currently set empirically and designed for future tuning based on feedback or supervised learning.

While the framework includes provisions for cosine similarity and embedding-based methods, they are reserved for future enhancements. This design ensures scalability, interpretability, and ease of integration into an evolving matchmaking platform.

B. Persona-Based Filtering via Clustering

To go beyond surface-level matching, our system builds meaningful user segments—called personas—using clustering algorithms like K-Means. We analyze structured attributes such as age, height, lifestyle choices (diet, smoking, drinking), and behavioral indicators (engagement, activity scores). These clusters help categorize users into groups like adventurous extroverts, career-focused introverts, or creative thinkers.

By assigning each user to a cluster, we enable persona-driven filtering—where match recommendations prioritize users with complementary or compatible behavioral traits. This ensures a more personalized and emotionally aligned dating experience, unlike traditional systems that match users solely by location or photos.

The clustering results also power features like tailored onboarding flows, niche community discovery, and refined recommendation logic based on evolving user behavior.

C. Mood-aware Chat and Story Sharing

Sammy Swipe enhances user interaction with a mood-aware chat system and integrated story sharing feature. The chat experience is elevated through real-time sentiment analysis, allowing the system to detect emotional tone and reflect it in chat themes or emoji suggestions. This helps users better understand each other's emotional state, making conversations more empathetic and engaging.

D. Social Media Integration

To make profiles more vibrant and authentic, Sammy Swipe allows users to connect their social media accounts, including Spotify, Instagram, and Facebook. With user consent, the app imports publicly available data like top songs, posts, hashtags, and profile aesthetics to enrich user profiles.

E. Real-Time Fake Profile Detection

The fraud detection system leverages Isolation Forest, an unsupervised anomaly detection algorithm well-suited for identifying rare or suspicious user behavior. It works by isolating data points that deviate significantly from the norm using randomly generated decision trees.

User profiles are represented using structured features such as bio length, interest count, login and message frequency, profile change history, and behavioral indicators like response time or suspicious login attempts. These features are standardized using feature scaling (StandardScaler) to ensure uniformity across attributes.

The model flags users as potential frauds if their behavior significantly differs from the general user base. A label of -1 indicates an anomaly, while 1 suggests normal activity. The use of Isolation Forest enables real-time fraud detection with minimal supervision and high interpretability.

XXI. System Screenshots

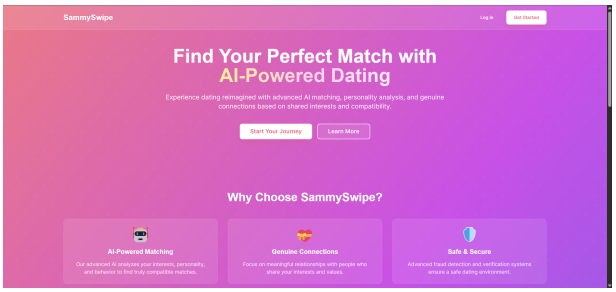


Fig. 21. SammySwipe landing page

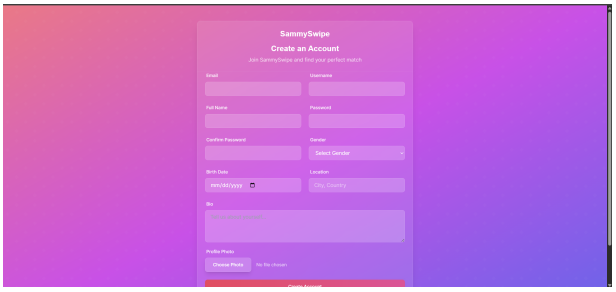


Fig. 22. SammySwipe profile creation page

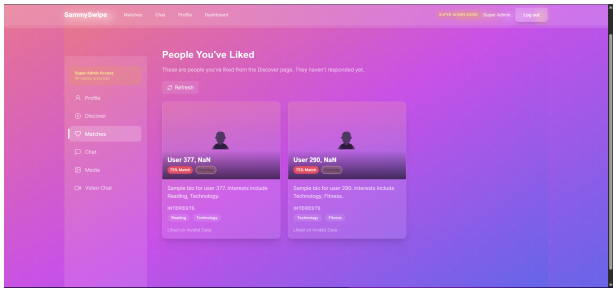


Fig. 23. SammySwipe matches page

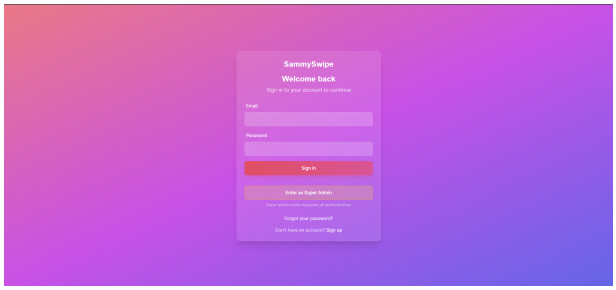


Fig. 24. SammySwipe login page

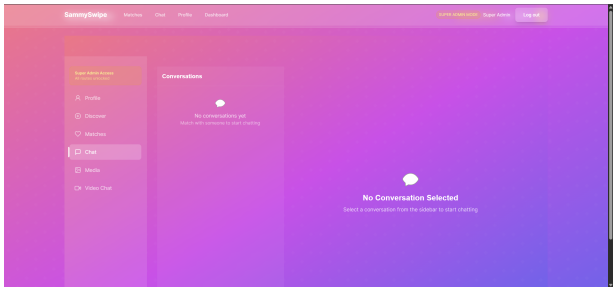


Fig. 25. SammySwipe chats page

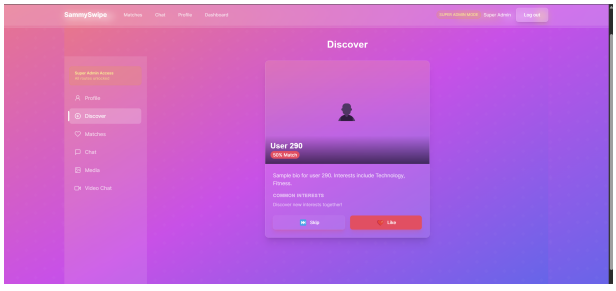


Fig. 26. SammySwipe discover page

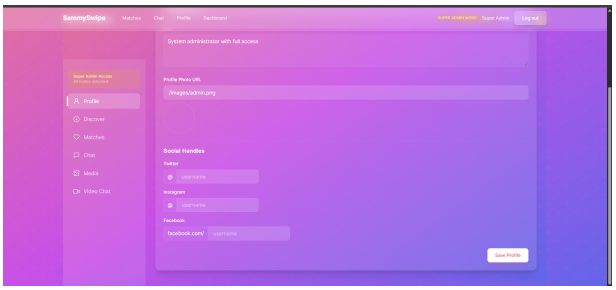


Fig. 27. SammySwipe profiles page

XXII. Advanced User Persona Development

Users are segmented via K-means clustering and profiled through NLP on bios and interactions. This dynamic segmentation adapts to user behavior and enhances match accuracy.

XXIII. Social Media Integration

Instagram, Spotify, and Facebook enrich profiles. NLP and image classification ensure depth and authenticity while maintaining user privacy via OAuth.

XXIV. Fake Profile Detection System

The fraud detection system uses Isolation Forest to flag suspicious user behavior based on anomalies in login patterns, bio length, and activity metrics. Features are scaled for consistency, enabling real-time, unsupervised detection with minimal overhead.

XXV. Implementation Workflow

1) Onboarding & Data Preprocessing
2) Persona Creation via Clustering
3) Graph Construction
4) Match Recommendation Engine
5) Fake Profile Detection & Feedback

XXVI. Evaluation Framework

- Match Quality: Engagement metrics
- Fake Detection Accuracy: F1-score, precision, recall
- Recommendation Performance: A/B testing
- User Satisfaction: Surveys

XXVII. Challenges and Future Scope

1) Cold Start Problem: Onboarding quizzes and inferred data can mitigate early data sparsity.

2) Privacy Concerns: Enhanced dashboards and anonymization will address data-sharing concerns.

3) Model Scalability: Real-time graph ops will be optimized via distributed computation.

4) Bias in Matching: Fairness audits and ethical AI design will ensure inclusivity.

5) Expansion to diverse Audience: Support for LGBTQ+ matching and language localization is planned.

Area	Current Gap	Planned Improvement
Privacy	Plain-text interest storage	AES-GCM field encryption (Q3)
Diversity	Binary gender only	Expand to spectrum + pronouns (Q2)
Cold-start	Sparse data for new users	Zero-shot interest inference via LLM embeddings (Q4)
Mobile	PWA only	Native iOS & Android apps (early 2026)

TABLE V
Planned Improvements to Address Current Gaps

XXVIII. Conclusion

Sammy Swipe redefines modern dating by putting real compatibility before casual swipes. Powered by machine learning, emotional intelligence, and graph-based insights, it helps users connect on a deeper level.
With smart persona modeling, social data integration, and strong safeguards against fake profiles, it ensures both meaningful matches and user safety. More than just a dating app, Sammy Swipe sets a benchmark for ethical, intelligent, and human-centered matchmaking in the digital age.

References

[1] Neo4j Graph Documentation. https://neo4j.com/docs/
[2] "Scikit-learn: Machine Learning in Python," JMLR, 2011.
[3] Mohammad and Turney, "Emotion Lexicon," ACL, 2013.
[4] Meta, "Instagram Graph API." https://developers.facebook.com/docs/instagram-api
[5] D. Hardt, "OAuth 2.0 Authorization Framework," RFC 6749, 2012.
[6] Mitchell et al., "Model Cards for Model Reporting," FAT, 2019.