RAJALAKSHMI
ENGINEERING
COLLEGE

Department of Artificial Intelligence and Data Science

A Big Data–Driven Customer Churn Prediction and Analysis System for Telecom Industry Using Databricks and PySpark YOLO

Mrs.SELVARANI
ASST.PROFESSOR

ASIYA BANU -231801015
DHARSHANA S-231801032
HARITHA M - 231801051

Problem Statement and Motivation

Conventional customer analysis methods in the telecom industry rely heavily on manual reports and static spreadsheets, which are **time-consuming, error-prone**, and unable to process massive volumes of customer data efficiently. Traditional approaches often fail to uncover complex patterns behind customer churn, **leading to inaccurate predictions** and delayed retention actions. Moreover, inconsistent data sources and missing information further reduce decision-making accuracy. Hence, there is a need for a scalable, data-driven churn prediction system using **Databricks and PySpark** that can handle large datasets, perform automated preprocessing, and **generate actionable insights** through interactive dashboards to help telecom providers proactively retain customers and reduce revenue loss.

Objectives

- □ To develop a scalable, end-to-end data pipeline using **Databricks and PySpark** for telecom customer churn analysis.
- □ To perform efficient **data ingestion, cleaning, and preprocessing** including handling missing values, encoding categorical data, and transforming numerical attributes
- □ To **analyze customer behavior** through interactive visualizations and identify key factors influencing churn such as contract type, payment method, and service usage.
- □ To build and evaluate **predictive machine learning models** capable of accurately identifying customers likely to churn.
- □ To deploy the analytical results through **Databricks SQL dashboards** for easy access by business and management teams to support data-driven retention strategies.

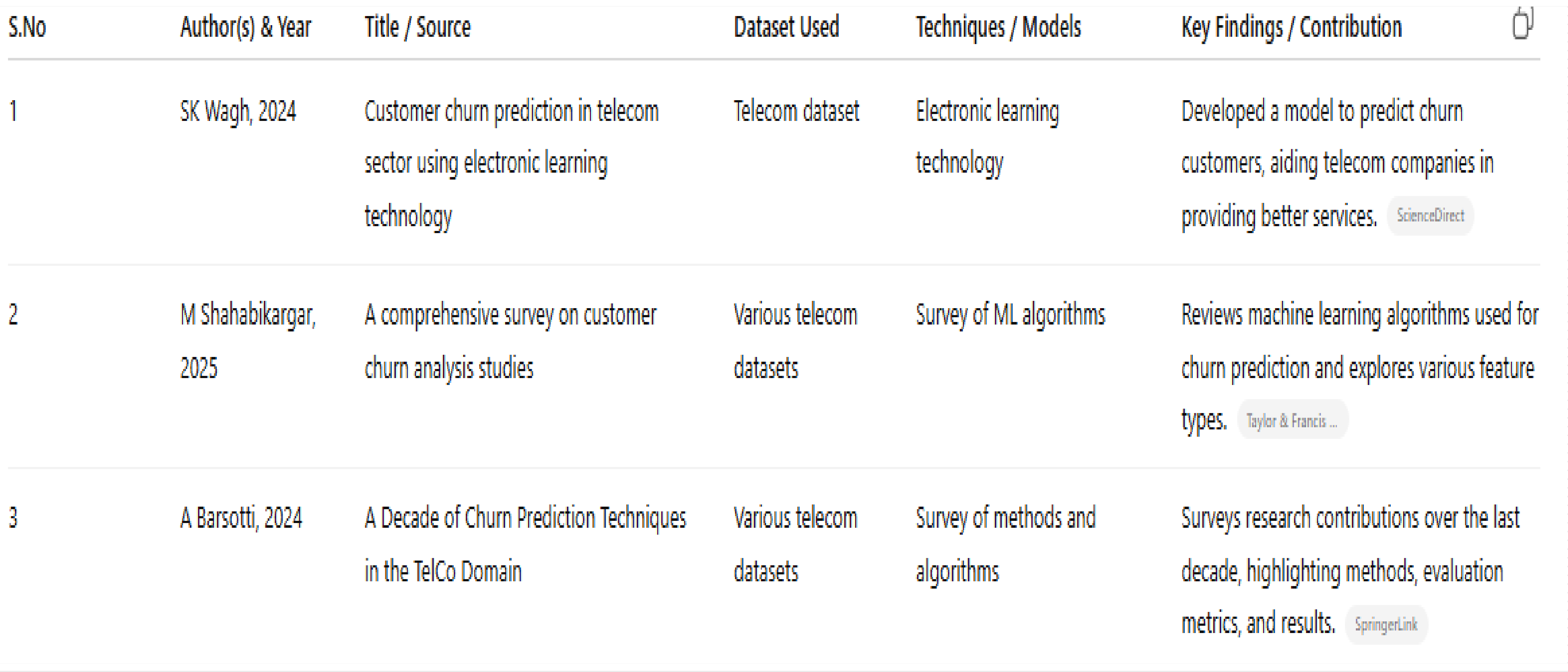
Existing System

S.no	Model/Technique Used	Author(s)	Description	Outcome
1.	Logistic Regression	IBM Sample Model (Telco Churn Dataset)	A traditional statistical method that models churn probability based on customer features such as tenure, monthly charges, and contract type.	Simple and interpretable but performs poorly on complex, non-linear relationships in large datasets.
2.	Decision Tree Classifier	Quinlan (1986)	Uses hierarchical splits of customer data based on feature importance to predict churn.	Provides better accuracy than Logistic Regression but prone to overfitting and limited scalability.
3.	Random Forest	Breiman (2001)	An ensemble learning technique combining multiple decision trees to improve generalization and reduce overfitting.	Achieves good accuracy on moderate datasets but computationally expensive for large-scale telecom data

Abstract

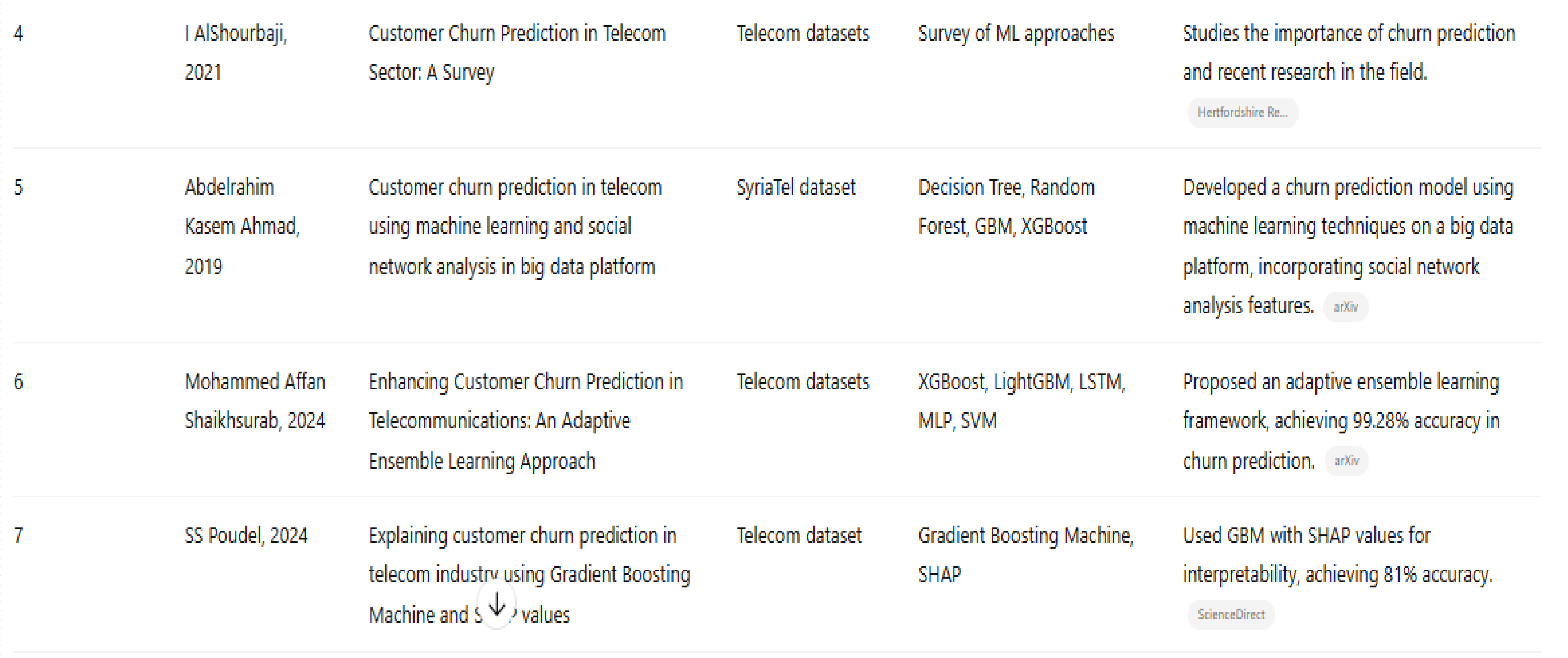
Customer churn poses a major challenge to telecom companies, directly affecting profitability and growth. Traditional manual analysis methods are slow, fragmented, and incapable of handling the massive volume of customer data generated daily. This project presents a **Big Data–driven customer churn prediction system** built using **Databricks and PySpark**. The system automates data ingestion, preprocessing, and analysis of the Telco customer dataset to uncover key churn indicators such as contract duration, service usage, and billing methods. By leveraging **PySpark’s distributed processing** and **Databricks SQL interactive dashboards**, the project delivers real-time insights and predictive analytics to help identify at-risk customers. Through this data-driven approach, telecom providers can enhance retention strategies, improve customer satisfaction, and reduce overall churn rates.

Literature Survey



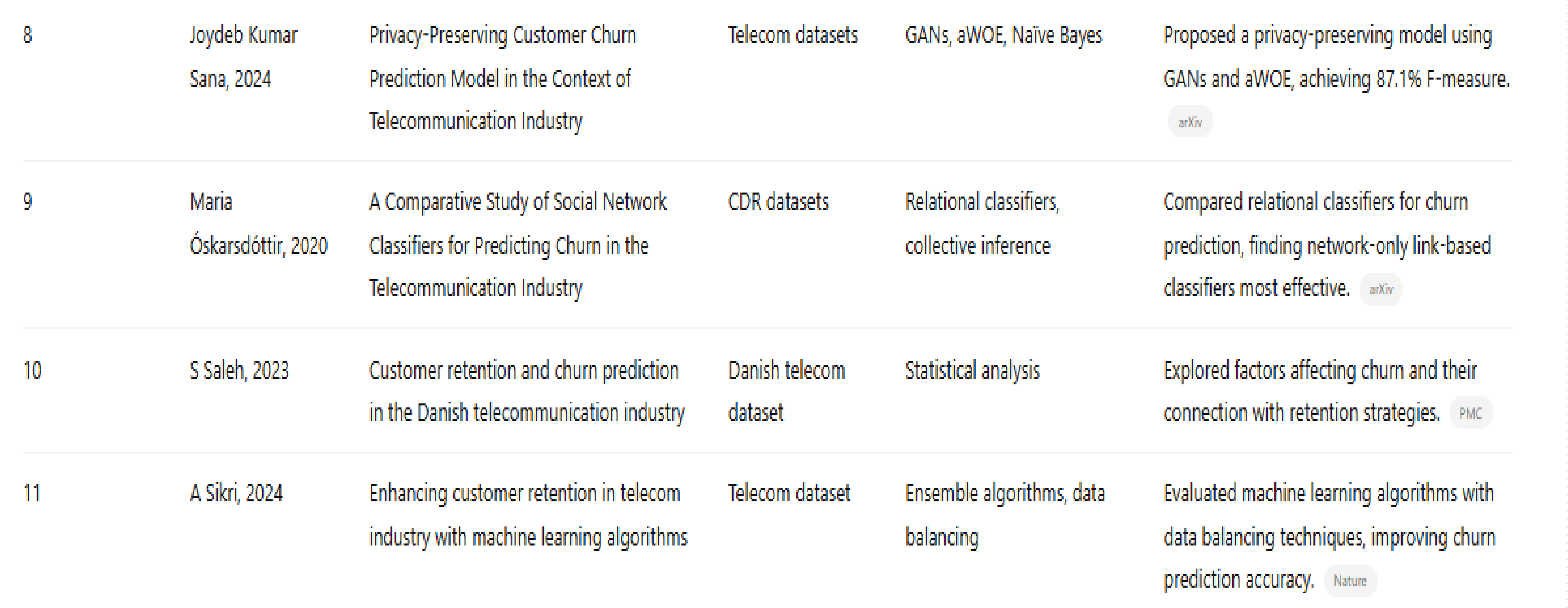
S.No	Author(s) & Year	Title / Source	Dataset Used	Techniques / Models	Key Findings / Contribution
1	SK Wagh, 2024	Customer churn prediction in telecom sector using electronic learning technology	Telecom dataset	Electronic learning technology	Developed a model to predict churn customers, aiding telecom companies in providing better services. ScienceDirect
2	M Shahabikargar, 2025	A comprehensive survey on customer churn analysis studies	Various telecom datasets	Survey of ML algorithms	Reviews machine learning algorithms used for churn prediction and explores various feature types. Taylor & Francis ...
3	A Barsotti, 2024	A Decade of Churn Prediction Techniques in the TelCo Domain	Various telecom datasets	Survey of methods and algorithms	Surveys research contributions over the last decade, highlighting methods, evaluation metrics, and results. SpringerLink

Literature Survey



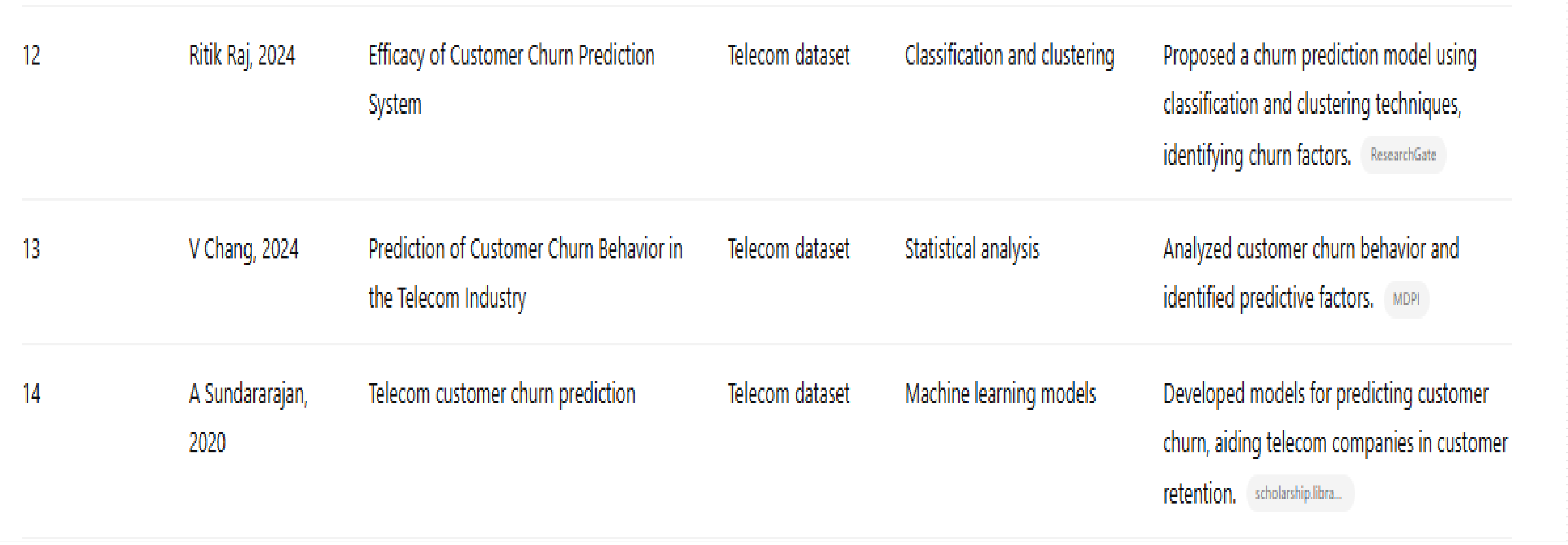
4	I AlShourbaji, 2021	Customer Churn Prediction in Telecom Sector: A Survey	Telecom datasets	Survey of ML approaches	Studies the importance of churn prediction and recent research in the field. Hertfordshire Re...
5	Abdelrahim Kasem Ahmad, 2019	Customer churn prediction in telecom using machine learning and social network analysis in big data platform	SyriaTel dataset	Decision Tree, Random Forest, GBM, XGBoost	Developed a churn prediction model using machine learning techniques on a big data platform, incorporating social network analysis features. arXiv
6	Mohammed Affan Shaikhsurab, 2024	Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach	Telecom datasets	XGBoost, LightGBM, LSTM, MLP, SVM	Proposed an adaptive ensemble learning framework, achieving 99.28% accuracy in churn prediction. arXiv
7	SS Poudel, 2024	Explaining customer churn prediction in telecom industry using Gradient Boosting Machine and SHAP values	Telecom dataset	Gradient Boosting Machine, SHAP	Used GBM with SHAP values for interpretability, achieving 81% accuracy. ScienceDirect

Literature Survey



8	Joydeb Kumar Sana, 2024	Privacy-Preserving Customer Churn Prediction Model in the Context of Telecommunication Industry	Telecom datasets	GANs, aWOE, Naïve Bayes	Proposed a privacy-preserving model using GANs and aWOE, achieving 87.1% F-measure. arXiv
9	Maria Óskarsdóttir, 2020	A Comparative Study of Social Network Classifiers for Predicting Churn in the Telecommunication Industry	CDR datasets	Relational classifiers, collective inference	Compared relational classifiers for churn prediction, finding network-only link-based classifiers most effective. arXiv
10	S Saleh, 2023	Customer retention and churn prediction in the Danish telecommunication industry	Danish telecom dataset	Statistical analysis	Explored factors affecting churn and their connection with retention strategies. PMC
11	A Sikri, 2024	Enhancing customer retention in telecom industry with machine learning algorithms	Telecom dataset	Ensemble algorithms, data balancing	Evaluated machine learning algorithms with data balancing techniques, improving churn prediction accuracy. Nature

Literature Survey



12	Ritik Raj, 2024	Efficacy of Customer Churn Prediction System	Telecom dataset	Classification and clustering	Proposed a churn prediction model using classification and clustering techniques, identifying churn factors. ResearchGate
13	V Chang, 2024	Prediction of Customer Churn Behavior in the Telecom Industry	Telecom dataset	Statistical analysis	Analyzed customer churn behavior and identified predictive factors. MDPI
14	A Sundararajan, 2020	Telecom customer churn prediction	Telecom dataset	Machine learning models	Developed models for predicting customer churn, aiding telecom companies in customer retention. scholarship.libra...

INTRODUCTION (SRS)

Title: Big Data–Driven Customer Churn Prediction and Analysis System for Telecom Industry Using Databricks and PySpark

Purpose:

- To predict customer churn accurately using large-scale telecom datasets
- To identify key behavioral and demographic factors contributing to customer attrition
- To enable telecom providers to make data-driven retention and marketing decisions.
- To provide real-time insights and visualizations through Databricks SQL dashboards for management and business teams.

Scope:

- Phase 1: Data ingestion and preprocessing of the Telco Customer Churn dataset using PySpark on Databricks; feature engineering and initial model training using Spark Mllib.
- Phase 2: Deployment of refined models with automated retraining, Mlflow experiment tracking and interactive dashboards in Databricks SQL for continuous churn monitoring and business insights.

Users:

- Telecom Marketing Teams, Customer Relationship Managers, Data Analysts and Data Scientisrs, Business Decision Makers

OVERALL DESCRIPTION

System Layers:

- **Frontend** → Databricks SQL **Interactive Dashboards** for churn visualization and key performance metrics.
- **Backend** → **PySpark on Databricks** for large-scale data ingestion, preprocessing, and machine learning model training
- **Database** → **Delta Lake / Parquet Storage** for maintaining cleaned datasets, feature tables, and model outputs.

Features:

Automated Data Processing: Cleans, transforms, and prepares large telecom datasets efficiently using PySpark.
Churn Prediction Models: Uses machine learning algorithms (Logistic Regression, Random Forest, Gradient Boosting) to identify customers likely to churn.
Insightful Dashboards: Visualizes churn distribution, service usage, and contract patterns through Databricks SQL.
Scalable & Reproducible Pipeline: Supports distributed data processing and model versioning with MLflow.
Actionable Intelligence: Helps telecom teams design retention campaigns based on churn probability and customer behavior patterns.

SYSTEM FEATURES

- **Data Ingestion** → Imports the Telco Customer Churn dataset into Databricks using PySpark, supporting large-scale and distributed data processing.
- **Data Preprocessing Module** → Handles missing values, data type conversions, encoding of categorical variables, and feature engineering for churn modeling.
- **Churn Prediction Model** → Trains and evaluates machine learning models such as Logistic Regression, Random Forest, and Gradient Boosting using Spark MLlib.
- **Visualization Dashboard** → Uses Databricks SQL Dashboards to display churn distribution, customer demographics, contract types, and payment method patterns.
- **Model Evaluation & Tracking** → Integrates MLflow for experiment tracking, accuracy comparison, and version control of models.
- **Data Storage** → Stores raw, processed, and model output data in Delta Lake / Parquet files, ensuring scalability and data consistency for future analysis.

ALGORITHMS

- **Machine Learning Models**: Logistic Regression, Random Forest, Gradient Boosting (XGBoost/LightGBM on Spark).
- Predicts **churn probability** for each customer based on demographic, contract, and service usage features.
- Confidence scores indicate **likelihood of churn** for targeted retention actions.

EXTERNAL INTERFACES

Prediction & Alerts:

- Real-time churn predictions with **probability scores** for each customer.
- Export **high-risk customer** lists to CRM/marketing systems.
- Interactive **Databricks SQL dashboards** for visualizing churn trends, cohorts, and KPIs.

Scalability & Storage:

- PySpark distributed processing for **large-scale telecom datasets**.
- Delta Lake / Parquet storage for **cleaned datasets**, features, and model outputs.
- Supports **automated pipeline** updates and scalable model retraining.

SYSTEM REQUIREMENTS

Hardware Requirements:
Minimum: 4 GB RAM, Dual-core CPU, 2 Mbps Internet, Camera Module
Recommended: 8 GB RAM, Quad-core CPU, SSD Storage, Stable Broadband, 5 MP+ CCTV/IP Camera

Software Requirements:
OS: Windows, Linux, or macOS
Python 3.10 or above, Flask or Django, PySpark, Pandas, NumPy, Matplotlib, Scikit-learn
Databricks Workspace with SQL Dashboard enabled
Database: PostgreSQL, MySQL, or MongoDB
Browser: Google Chrome (latest version) with extensions enabled

NON FUNCTIONAL REQUIREMENTS

- **Performance:** Model should generate churn predictions in under 2 seconds for batch or real-time requests.
- **Security:** All customer data should be encrypted and handled securely with restricted database access.
- **Scalability:** The system should efficiently handle millions of customer records and support distributed processing via Databricks clusters.
- **Portability:** The solution should run on multiple platforms — Windows, Linux, or cloud environments (Databricks, AWS, Azure).
- **Maintainability:** Modular architecture with separate components for data ingestion, preprocessing, modeling, and visualization for easy updates.

FUTURE SCOPE

- **Real-Time Churn Prediction:** Integrate live data streams to predict churn in real-time as customer behavior changes.
- **Advanced Modeling:** Implement ensemble and deep learning models (e.g., XGBoost, Neural Networks) to improve prediction accuracy.
- **Automated Insights:** Generate personalized retention recommendations based on churn risk scores.
- **Integration with CRM Systems:** Connect the churn prediction dashboard with customer management tools for proactive outreach.
- **Enhanced Visualization:** Add more interactive dashboards and predictive trend analysis for business decision-making.
- **Maintainability:** Modular architecture with separate components for data ingestion, preprocessing, modeling, and visualization for easy updates.

Thank You