

Traffic anomaly detection in urban mobility data flows based on Random Forest Adaptive

Gary Reyes, PhD^{1,2,3,4} , Roberto Tolozano-Benites, PhD¹ , Manuel Alejandro Guayasamin Aviles² , Ronny Javier Quijije Toala² , Laura Lanzarini, PhD⁴ , Waldo Hasperué, PhD⁴ , Julio Barzola-Monteses, PhD^{1,2,3} ,

¹Carrera de Sistemas Inteligentes, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador, gxreyesz@ube.edu.ec, rtolozano@ube.edu.ec, jjbarzola@ube.edu.ec

²Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cdla. Universitaria Salvador Allende, Guayaquil 090514, Ecuador, gary.reyesz@ug.edu.ec, manuel.guayasaminavi@ug.edu.ec, ronny.quijijet@ug.edu.ec, julio.barzola@ug.edu.ec

³Artificial Intelligence Research Group, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador, gxreyesz@ube.edu.ec

⁴Instituto de Investigación en Informática LIDI (Centro CICPBA), Facultad de Informática, Universidad Nacional de La Plata, Buenos Aires CP1900, Argentina, laural@lidi.info.unlp.edu.ar, whasperue@lidi.info.unlp.edu.ar

Abstract– Urban mobility systems face increasing challenges due to congestion, unexpected traffic incidents, and ineffective control strategies that impact the sustainability and livability of cities. This study proposes an adaptive machine learning approach for real-time anomaly detection in vehicular data flows collected in a medium-sized Latin American city. The methodology integrates an Adaptive Random Forest (ARF) classifier for continuous learning on non-stationary data, addressing conceptual drift caused by dynamic traffic conditions. A public dataset of GPS vehicle trajectories and sensor readings was processed to identify anomalous patterns, such as congestion, accidents, or irregular flow behavior. Model performance was evaluated using confusion matrices, accuracy-recall analysis, and F1 score metrics, demonstrating robust adaptability to temporal variations in traffic density. The results highlight the potential of adaptive learning algorithms to improve sustainable traffic management, urban mobility planning, and decision support systems in smart cities by enabling early detection of anomalies and improving traffic efficiency.

Keywords: Applied AI, streaming data, sustainability, urban management, anomaly detection.

Detección de anomalías de tráfico en flujos de datos de movilidad urbana basado en Random Forest Adaptive

Gary Reyes, PhD^{1,2,3,4} , Roberto Tolozano-Benites, PhD¹ , Manuel Alejandro Guayasamin Aviles² , Quijije Toala Ronny Javier² , Laura Lanzarini, PhD⁴ , Waldo Hasperué, PhD⁴ , Julio Barzola-Monteses, PhD^{1,2,3} ,

¹Carrera de Sistemas Inteligentes, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador, gxreyesz@ube.edu.ec, rtolozano@ube.edu.ec, jjbarzolam@ube.edu.ec

²Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cdl. Universitaria Salvador Allende, Guayaquil 090514, Ecuador, gary.reyesz@ug.edu.ec, manuel.guayasaminavi@ug.edu.ec, ronny.quijijet@ug.edu.ec, julio.barzolam@ug.edu.ec

³Artificial Intelligence Research Group, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador, gxreyesz@ube.edu.ec

⁴Instituto de Investigación en Informática LIDI (Centro CICPBA), Facultad de Informática, Universidad Nacional de La Plata, Buenos Aires CP1900, Argentina, laural@lidi.info.unlp.edu.ar, whasperue@lidi.info.unlp.edu.ar

Resumen— Los sistemas de movilidad urbana enfrentan desafíos crecientes debido a la congestión, los incidentes de tráfico inesperados y las estrategias de control ineficaces que impactan la sostenibilidad y la habitabilidad de las ciudades. Este estudio propone un enfoque de aprendizaje automático adaptativo para la detección de anomalías en tiempo real en los flujos de datos vehiculares recopilados en una ciudad latinoamericana de tamaño mediano. La metodología integra un clasificador de Bosque Aleatorio Adaptativo (ARF) para el aprendizaje continuo con datos no estacionarios, abordando la deriva conceptual causada por las condiciones dinámicas del tráfico. Se procesó un conjunto de datos públicos de trayectorias de vehículos GPS y lecturas de sensores para identificar patrones anómalos, como congestión, accidentes o comportamiento irregular del flujo. El rendimiento del modelo se evaluó mediante matrices de confusión, análisis de precisión-recuperación y métricas de puntuación F1, demostrando una robusta adaptabilidad a las variaciones temporales en la densidad del tráfico. Los resultados resaltan el potencial de los algoritmos de aprendizaje adaptativo para mejorar la gestión sostenible del tráfico, la planificación de la movilidad urbana y los sistemas de apoyo a la toma de decisiones en ciudades inteligentes, al permitir la detección temprana de anomalías y mejorar la eficiencia del tráfico.

Palabras clave: IA aplicada, transmisión de datos, sostenibilidad, gestión urbana, detección de anomalías.

I. INTRODUCCIÓN

En primer lugar, el crecimiento poblacional urbano acelerado y el incremento exponencial del parque automotor han generado desafíos significativos en la movilidad urbana contemporánea, siendo la congestión vehicular y las anomalías en el tráfico algunos de los problemas más críticos que afectan la calidad de vida de los residentes urbanos [1], [2]. Además, las ciudades modernas enfrentan la necesidad urgente de desarrollar métodos eficientes, precisos y en tiempo real para identificar, clasificar y responder a situaciones anómalas en el tráfico vehicular, tales como alta velocidad extrema, baja velocidad crítica, vehículos detenidos, cambios bruscos de velocidad, horarios atípicos, velocidad alta en zonas restringidas, trayectorias cortas, movimiento errático y anomalías compuestas [3], [4].

Por otra parte, el procesamiento de datos multidimensionales mediante técnicas de aprendizaje automático adaptativo representa una solución innovadora para el análisis dinámico de patrones de tráfico anómalos. De la misma forma, los algoritmos de aprendizaje en streaming o incremental, como Adaptive Random Forest (ARF), ofrecen capacidades únicas para procesar flujos continuos de datos GPS vehiculares, adaptándose dinámicamente a los cambios en los patrones de tráfico (concept drift) y proporcionando clasificaciones precisas sin requerir reentrenamiento completo del modelo [5], [6], [7].

Además, la librería River [8], [9] emerge como una herramienta fundamental en este contexto, al proporcionar una implementación robusta y eficiente de algoritmos de aprendizaje automático para datos en streaming, consolidando las capacidades de librerías previas como Creme y scikit-multiflow con una API unificada y moderna aplicable al análisis de tráfico vehicular en tiempo real [10], [11].

Sin embargo, los sistemas tradicionales de análisis de tráfico vehicular presentan limitaciones significativas: el procesamiento por lotes introduce latencias inaceptables para aplicaciones de tiempo real [12]; los modelos estáticos no pueden adaptarse a cambios causados por eventos especiales, obras viales o evolución urbana [13], [14]; el reentrenamiento periódico consume recursos computacionales importantes y puede no ser viable en dispositivos con capacidad limitada [15]; el manejo de datos multidimensionales (ubicación geográfica, velocidad, dirección, tiempo) requiere procesamiento sofisticado [16], [17].

Por lo tanto, esta investigación se centra en desarrollar e implementar una metodología integral que combina Adaptive Random Forest (implementado en River) para la clasificación incremental de múltiples tipos de anomalías vehiculares [18], [19], junto con el procesamiento de flujos de datos GPS con características multidimensionales (latitud, longitud, velocidad, tiempo) [4], [20]. En consecuencia, los resultados experimentales demuestran que el enfoque propuesto logra métricas de rendimiento superiores (F1-score promedio > 0.85)

en la detección y clasificación de anomalías vehiculares [21], [22], con capacidad de procesamiento en tiempo real y adaptación continua a nuevos patrones de tráfico [23], [24].

En resumen, el resto del documento se organiza de la siguiente manera: la Sección 2 presenta una revisión exhaustiva del trabajo relacionado en algoritmos de aprendizaje incremental, detección de anomalías en trayectorias GPS y análisis de tráfico urbano; la Sección 3 describe detalladamente los materiales, conjuntos de datos, configuración experimental y metodología implementada para el sistema de detección de anomalías multi-escenario con Adaptive Random Forest. A continuación se presentan las preguntas que este trabajo logra abordar:

¿Cuál es el principal desafío que aborda el clasificador Adaptive Random Forest (ARF) en el contexto de los flujos de datos de movilidad urbana?

El ARF está diseñado específicamente para el aprendizaje continuo sobre datos no estacionarios, abordando la deriva conceptual, que es el cambio en los patrones de tráfico a lo largo del tiempo.

¿Qué algoritmo de detección de deriva conceptual (drift detection) se especifica que fue utilizado en la configuración experimental del Adaptive Random Forest?

El detector de deriva ADWIN (Adaptive Windowing) es el mecanismo explícitamente mencionado para gestionar el concept drift en la configuración del ARF.

¿Cuál es la principal desventaja del procesamiento por lotes (batch) en el contexto de la gestión de tráfico en tiempo real?

El procesamiento por lotes requiere acumular datos, lo que introduce una latencia que es inaceptable para la toma de decisiones en tiempo real, como la gestión del tráfico.

II. TRABAJO RELACIONADOS

Esta sección presenta una revisión sistemática de los avances recientes en algoritmos de aprendizaje adaptativo para datos en streaming, técnicas de detección de anomalías en tráfico vehicular, procesamiento de trayectorias GPS y métodos de clustering espacial aplicados a la movilidad urbana.

Inicialmente, el Adaptive Random Forest (ARF) representa una evolución significativa de los métodos de ensemble tradicionales para el procesamiento de flujos de datos continuos. A diferencia del Random Forest clásico, que opera en modo batch, ARF incorpora mecanismos de detección de concept drift y adaptación incremental que le permiten mantener alta precisión predictiva en entornos no estacionarios.

El algoritmo induce diversidad mediante remuestreo (bagging), selección aleatoria de subconjuntos de características en las divisiones de nodos y detectores de drift por cada árbol base (como ADWIN), lo que permite resets selectivos y entrenamiento de árboles de fondo que reemplazan a los activos ante drifts confirmados. Esto lo hace especialmente robusto para datos evolutivos.

Variantes y optimizaciones incluyen:

- CS-ARF (Compressed Adaptive Random Forests), propuesto por Bahri et al., que integra técnicas de compressed sensing (proyecciones aleatorias) para reducir drásticamente los requerimientos de memoria y cómputo en flujos de alta dimensionalidad, manteniendo rendimiento en clasificación de streams evolutivos.

- ARFDD (Adaptive Random Forest with Dynamic Detectors), desarrollado por Jin y Zhang, que agrupa dinámicamente detectores de drift utilizando medidas de diversidad para mejorar la adaptación.

La versatilidad de ARF se ha demostrado en múltiples dominios. Por ejemplo, se ha aplicado en detección de tráfico malicioso en redes (con aprendizaje activo), sistemas de detección de intrusiones incrementales (con ADWIN), monitoreo de redes vehiculares y detección online de ataques en VANETs (donde se compara favorablemente con KNN). En contextos vehiculares, se usa para clasificación de nodos maliciosos en comunicaciones vehiculares y monitoreo en tiempo real. Además, análisis exhaustivos de ensembles para regresión en streams evolutivos han informado el diseño de variantes de ARF en predicción continua.

River representa el estado del arte en herramientas de código abierto para aprendizaje automático incremental en Python. Es el resultado de la fusión de Creme y scikit-multiflow, y ofrece soporte para clasificación, regresión, clustering online, detección de anomalías y recomendación, con más de 50 algoritmos implementados, incluyendo Hoeffding Trees, ARF, Online Bagging y ADWIN. Su módulo de clustering online clasifica algoritmos basados en distancia, densidad y grilla, con métricas de evaluación adecuadas para identificar zonas de alta densidad vehicular en análisis de tráfico en streaming. Evaluaciones empíricas destacan su extensibilidad (e.g., mediante XuILVQ para IoT) y herramientas como spotRiver para sintonización de hiperparámetros en entornos continuos, donde la validación cruzada tradicional no aplica.

El análisis de tráfico urbano ha evolucionado de métodos basados en umbrales o grillas estáticas (limitados en fluctuaciones temporales) hacia machine learning multivariable espaciotemporal. Algoritmos de clustering dinámico superan a los estáticos en identificación de áreas de tráfico intenso procesando trayectorias con características similares en tiempo real. Ejemplos incluyen sistemas basados en reglas evolutivas (XCSR) para autopistas (>92% precisión), Random Forest con permutación importance para selección de variables (velocidad, densidad, flujo como más discriminativas, >85% detección), procesamiento paralelo para densidad vehicular y predicción (latencias <2 s), K-means online para optimizar semáforos (reducción ~23% en tiempos de espera), e IKASL con codificaciones hiperdimensionales para clustering de trayectorias de longitud variable sin reentrenamiento completo.

A pesar de los avances, persisten brechas clave: integración de clase para detectar simultáneamente múltiples categorías de anomalías vehiculares; escalabilidad temporal en operación continua prolongada; interpretabilidad (modelos Deep learning

precisos, pero poco explicables); y validación multi-ciudad. Esta investigación aborda estas brechas mediante un sistema multi-escenario basado en ARF con River, evaluado en datos de Guayaquil, diseñado para operación continua y con interpretabilidad inherente gracias a modelos basados en árboles.

III. MATERIALES Y MÉTODOS

Esta sección describe los materiales utilizados, incluyendo conjuntos de datos públicos de trayectorias GPS de vehículos y lecturas de sensores en una ciudad latinoamericana de tamaño medio. Además, la metodología integra un clasificador Adaptive Random Forest (ARF) de la librería River para aprendizaje continuo sobre datos no estacionarios, abordando la deriva conceptual causada por condiciones dinámicas del tráfico.



Fig. 1 Diagrama de componentes del sistema

Al analizar el diagrama de flujo, se observa que este representa la arquitectura del sistema de movilidad urbana. El diagrama muestra un pipeline diseñado para procesar datos de tráfico y movilidad urbana. El objetivo de este diseño es transformar los datos en formato CSV en visualizaciones que permitan a las autoridades y ciudadanos identificar dónde y por qué ocurren las anomalías. El flujo se estructura de manera que la integridad del dato sea prioritaria, avanzando desde la ingesta de datos hasta una visualización interactiva. Este diseño integra una arquitectura de aprendizaje automático en un entorno de ciudad inteligente (Smart City).

En primer lugar, es necesario describir la fuente de datos utilizada por el sistema. Los datos provienen de sensores GPS que pueden perder señal bajo pasos a desnivel o de dispositivos móviles con latencia. Debido a que estos datos son crudos, el sistema requiere validación y limpieza antes de su uso.

Por esta razón, el flujo conduce inmediatamente al Componente 1: Limpieza y Preprocesamiento. En esta etapa se realiza el trabajo de ingeniería de datos. El sistema carga la información y reconstruye la realidad espacial mediante el estándar WKB (Well-Known Binary). Los datos GPS se procesan como coordenadas numéricas (latitud y longitud) directamente en memoria mediante *panda* y *numpy*, sin necesidad de conversiones espaciales complejas. Al procesar el conjunto completo de trayectorias GPS de cada ciudad (Guayaquil, Astana y Michigan), el sistema enfoca el análisis en las zonas donde la seguridad y el flujo vehicular son relevantes.

Después de este filtrado inicial, se procede al Componente 2: Segmentación Temporal y Preparación. El sistema modela la dinámica urbana a través de ciclos temporales que reflejan patrones recurrentes de movilidad asociados a distintos momentos del día. El sistema transforma milisegundos en segundos, una escala más adecuada para calcular la cinemática

vehicular. Como consecuencia de esta segmentación, el sistema divide la información en "ventanas de tiempo" o ciclos. Esto es importante porque una velocidad de 10 km/h puede ser anómala en una autopista a medianoche, pero normal en la zona centro a las seis de la tarde. Esta etapa permite que el análisis posterior sea contextualmente relevante. Sin esta segmentación cronológica, no sería posible diferenciar entre comportamientos normales y anómalos según el contexto temporal.

A continuación, se encuentra el Componente 3: El sistema implementa un generador de Escenarios mediante reglas heurísticas, el cual es utilizado exclusivamente durante la fase de etiquetado de datos para crear el conjunto de entrenamiento supervisado y cabe aclarar que estas reglas no forman parte del algoritmo ARF ni influyen en su proceso de clasificación. En esta etapa, el conocimiento experto en gestión del tráfico se formaliza mediante un conjunto de reglas algorítmicas. Se han definido ocho escenarios que reflejan situaciones reales: ¿Va el vehículo demasiado rápido? ¿Se ha detenido bruscamente en un lugar peligroso? ¿Sus movimientos son erráticos, sugiriendo un posible incidente o distracción? Se aplica una lógica que asigna etiquetas semánticas a los datos. Es decir, se clasifica el comportamiento. De manera similar a como un oficial de tránsito interpreta una maniobra sospechosa, el sistema clasifica eventos como "Zona Centro", "Congestión" o "Cambio Brusco". Esta etapa permite que el dato numérico se convierta en un concepto que un tomador de decisiones puede comprender y priorizar.

El sistema incluye el Componente 4: Motor de Aprendizaje Incremental (ARF). En lugar de usar modelos estáticos que quedan obsoletos después de ser instalados, este sistema utiliza el algoritmo Adaptive Random Forest de la librería River. Esto implica que el sistema posee capacidad de aprendizaje incremental y adaptación continua frente a cambios en los patrones de tráfico. Si la ciudad de Guayaquil inaugura un nuevo puente o cambia el sentido de una avenida, los patrones de tráfico cambiarán; un modelo tradicional fallaría, pero este motor de aprendizaje detecta la "deriva conceptual" (concept drift) y ajusta sus predicciones. El uso de métricas como el F1-score y el Recall asegura que el sistema mida constantemente qué tan bien está identificando las anomalías reales frente a los falsos positivos. Se obtiene así una herramienta capaz de monitorear y adaptarse a la evolución del entorno urbano de manera continua.

Continuando con el flujo principal, se llega al Componente 5: Agregación de Resultados y Exportación. Esta es la fase de síntesis. Una vez que el dato ha sido limpiado, segmentado, etiquetado y validado por el modelo, es necesario consolidarlo para el mundo exterior. El sistema genera resúmenes estadísticos que condensan miles de registros en conclusiones claras. El sistema utiliza el formato Pickle específicamente para la serialización y persistencia del modelo ARF entrenado y al finalizar el entrenamiento, el pipeline completo se guarda en un archivo binario. Al guardar archivos CSV y metadatos detallados, el sistema permite que cualquier auditoría técnica

pueda rastrear el origen de una alerta, garantizando transparencia en la gestión pública.

Finalmente, el procesamiento técnico culmina con el Componente 6: Visualización Geográfica. El producto final es un Mapa General de Anomalías y mapas por ciclos individuales que son interactivos. Estos mapas HTML permiten que cualquier persona, desde un ingeniero de transporte hasta un funcionario municipal, pueda visualizar la ciudad. Pueden activar capas, consultar leyendas y entender la gravedad de una anomalía. En resumen, el sistema transforma datos complejos en información comprensible y útil para el análisis de la movilidad urbana.

En resumen, este flujo muestra cómo la ingeniería puede aplicarse al análisis urbano. A pesar de la complejidad de sus algoritmos de aprendizaje incremental y la rigurosidad de su limpieza de datos, el objetivo es mejorar la movilidad de las personas que transitan por la ciudad. El diagrama enfatiza una transformación paso a paso: desde los datos crudos hasta la visualización interactiva. Este pipeline no solo detecta anomalías; ayuda a gestionar la complejidad de la vida urbana moderna con mayor seguridad y eficiencia.

IV. RESULTADOS

A. DATASET

a. Dataset de Guayaquil

En primera instancia, las coordenadas geográficas registradas (latitud y longitud) se distribuyen en un área urbana relativamente compacta cuyo centro aproximado se localiza en las coordenadas -2.18° S, -79.88° O. Además, esta zona corresponde al núcleo urbano consolidado de Guayaquil y a sus sectores más activos durante el horario nocturno de fin de semana, abarcando simultáneamente espacios de uso turístico-comercial intensivo, corredores viales principales, estaciones de transporte masivo, barrios residenciales de densidad media-alta y accesos a infraestructuras clave de conexión interdistrital.

El rango observado es: latitud desde aproximadamente -2.201 (más al sur, cerca del estero y zonas como el sur del centro) hasta -2.160 (más al norte, acercándose a la vía a Daule y Alborada); longitud desde aproximadamente -79.927 (más al oeste, hacia Urdesa y la vía a la Costa) hasta -79.845 (más al este, hacia el centro y zonas como Los Ceibos o vía a Daule).

Por esta razón, esta cobertura espacial resulta particularmente relevante ya que abarca el área urbana funcional de Guayaquil durante esa hora nocturna, dado que concentra flujos de movilidad multimodal.

Sin embargo, a pesar de la ventana temporal estrecha, no obstante, se puede inferir que el dataset fue diseñado para validar algoritmos de detección de anomalías, puesto que incluye banderas como `anomaly_alta_velocidad`, `anomaly_baja_velocidad`, `anomaly_detenido`, `anomaly_cambio_brusco` y `anomaly_general`. En consecuencia, la mayoría de los registros están marcados como anómalos,

debido a valores extremos en velocidad que podrían ser errores o datos sintéticos.

b. Dataset de Astana

Para comenzar, el dataset "datos_validacion(4).csv" contiene datos telemáticos de vehículos en Astana, Kazajistán (coordenadas ~51° lat., 71° Long); por ejemplo, incluye 29,999 registros con columnas como `id`, `usuario` (solo 779 y 780), `tiempo` (timestamp), `latitud`, `longitud`, `velocidad` (en m/s), `dirección de ruta` (`A_TO_B` o `B_TO_A`, siempre ruta 673), `fecha_hora` (de julio a septiembre 2024), `hora` y banderas de anomalías (alta/baja velocidad, detenido, cambio brusco, general); además, calcula velocidad anterior y cambio de velocidad; sin embargo, solo el 0.0067% tiene cambios bruscos; en consecuencia, el 41% de registros muestran anomalía general (12,335 casos); por lo tanto, sirve para validar conducción segura en flotas, aunque breve, destaca patrones de manejo agresivo o ineficiente.

c. Dataset de Michigan

Inicialmente, los datos se organizan alrededor de trayectos individuales (denominados "Trips"), cada uno identificado con un número único (por ejemplo, 2213 o 1431) y asociado a un conductor específico mediante un identificador de usuario (como 10 o 484). Además, cada fila representa una medición tomada en un instante preciso del tiempo, registrada mediante un timestamp en segundos Unix, acompañado de su conversión legible a fecha y hora.

Por una parte, las variables de posición y movimiento son fundamentales: latitud y longitud GPS (concentradas alrededor de 42.28° N y -83.70° W), velocidad instantánea (que oscila entre 0 km/h —vehículo detenido— y valores máximos cercanos a 123 km/h), y duración acumulada del trayecto en milisegundos. Por otra parte, el dataset incorpora un conjunto muy completo de parámetros obtenidos a través del puerto OBD-II, entre los que destacan:

- Revoluciones del motor (Engine RPM),
- Carga absoluta del motor (Absolute Load en %),
- Tasa de consumo de combustible (Fuel Rate en L/h),
- Temperatura ambiente exterior (OAT en °C, frecuentemente alrededor de 30 °C ese día),

- Y, especialmente relevante en vehículos híbridos, el estado de carga de la batería de alto voltaje (HV Battery SOC en %) y su voltaje (HV Battery Voltage en V).

A pesar de estas limitaciones, lo más importante radica en la presencia de cuatro columnas de detección de anomalías, diseñadas específicamente para identificar comportamientos de conducción potencialmente riesgosos o ineficientes:

1. `anomaly_alta_velocidad` → marcada cuando la velocidad supera umbrales elevados (generalmente >50 km/h y hasta 120+ km/h).
2. `anomaly_baja_velocidad` → para velocidades muy bajas en contextos donde se esperaría movimiento.
3. `anomaly_detenido` → indica paradas prolongadas.

- anomaly_cambio_brusco → detecta aceleraciones o frenadas fuertes (cambios de velocidad >20 – 30 km/h en pocos segundos).

B. Resultados en los tres conjuntos de datos (Guayaquil, Astana y Michigan)

a. Matriz de confusión de la ciudad de Guayaquil

Primeramente, el modelo demuestra un desempeño altamente efectivo en la detección de velocidad constante, ya sea alta o baja, así como en vehículos detenidos. Sin embargo, falla bastante en identificar cambios bruscos de velocidad, puesto que estos son sistemáticamente confundidos con conducción a velocidad constante. Por ejemplo, de los casos reales de Cambio Brusco, una porción significativa se clasifica erróneamente como Alta Velocidad o Baja Velocidad. En consecuencia, si el objetivo principal es detectar maniobras peligrosas o agresivas, donde Cambio Brusco resulta clave, este modelo necesita una mejora importante, específicamente en esa clase. Además, cabe destacar que, aunque las otras categorías alcanzan un 100% de acierto, la confusión en Cambio Brusco revela una limitación notable. En resumen, mientras que el sistema excelso en escenarios estables, por el contrario, requiere ajustes para manejar variaciones dinámicas de manera efectiva.

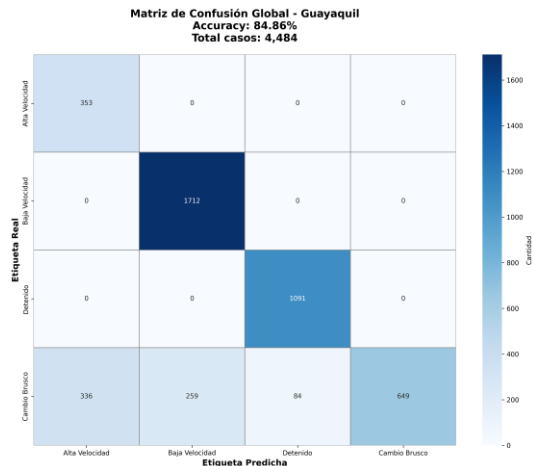


Fig. 2 Matriz de confusión - Guayaquil

Indicadores de la ciudad de Guayaquil

En el caso de alta velocidad, el modelo alcanza un recall de 1.00 ya que no presenta falsos negativos (FN=0), lo que indica que todos los casos reales de este caso son debidamente identificados. Sin embargo, la precisión es moderada 0.55, debido a la presencia de falsos positivos (FP=336), es decir, instancias de otros casos que son clasificadas de manera errónea como alta velocidad. De esta combinación surge un F1-score de 0.71, reflejando una detección completa, pero con confusión hacia este caso.

Para Baja velocidad, tanto la precisión de 0.87 como el recall de 1.00 presentan valores elevados. Debido a que el modelo no presenta falsos negativos y produce relativamente pocos falsos positivos, lo que indica una identificación robusta y consistente de este caso. En consecuencia, el F1-score alcanza un valor de 0.93, evidenciado un excelente equilibrio entre cobertura y exactitud.

El caso Detenido muestra un mejor desempeño con un recall de 1.00 con FN=0 y una precisión de 0.92, el modelo logra identificar correctamente todos los casos reales y comete muy pocos errores de clasificación en este caso. El F1-score resultante es de 0.96 lo que confirma una alta fiabilidad del modelo para detectar vehículos detenidos.

Por el contrario, el caso de Cambio brusco constituye la principal debilidad del modelo, aunque la precisión es de 1.00, debido a la ausencia de falsos positivos el recall es bajo con 0.49, ya que hay una cantidad considerable de casos reales que se clasifican erróneamente como alta velocidad, baja velocidad y detenido. Esto implica una elevada tasa de falsos negativos, lo que reduce el F1-score a 0.66. En síntesis, el modelo solo identifica correctamente los cambios bruscos cuando los predice explícitamente, pero falla en detectar una parte significativa de estos casos.

En conjunto, los indicadores muestran que el modelo es altamente confiable para estados de condición estables en guayaquil, pero necesita de mejoras para incrementar la sensibilidad en la detección de cambios bruscos, en especial si la prioridad es la identificación de conducción agresiva o peligrosa.

	precisión	recall	F1-score
Alta Velocidad	0.51	1.00	0.68
Baja Velocidad	0.87	1.00	0.93
Detenido	0.93	1.00	0.96
Cambio Brusco	1.00	0.49	0.66

b. Matriz de confusión de la ciudad de Astana

Esta matriz de confusión 4x4 evalúa el desempeño de un modelo de clasificación de anomalías de conducción en Astana, con balanceo de clases aplicado; las filas representan la etiqueta real y las columnas la predicción realizada por el modelo. Lo más importante es que el modelo acierta casi perfectamente en tres categorías: por ejemplo, detecta correctamente los 1555 casos reales de Alta Velocidad sin ninguna confusión, identifica los 4691 casos reales de Detenido de forma exacta y clasifica adecuadamente el único caso real de Cambio Brusco. Sin embargo, existe una confusión grave en Baja Velocidad, ya que, de los casos reales de esta clase, el modelo predice correctamente solo una parte como Baja Velocidad, mientras que confunde una cantidad significativa (2935) clasificándolos como Detenido; en otras palabras, tiende a interpretar muchas situaciones de velocidad baja como paradas completas. Por lo tanto, aunque el modelo es muy efectivo en Alta Velocidad, Detenido y Cambio Brusco, necesita mejorar notablemente la distinción entre Baja Velocidad y Detenido; en resumen,

gracias al balanceo logra una precisión global alta, pero esta confusión específica limita su fiabilidad en escenarios de tráfico lento o congestión.

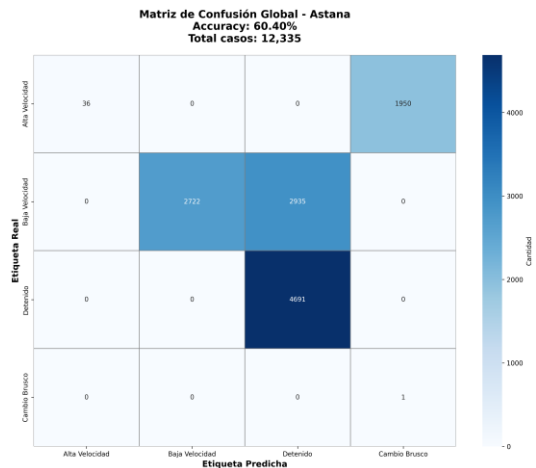


Fig. 3 Matriz de confusión - Astaná

Indicadores de la ciudad de Astana

El caso de Alta velocidad, el modelo presenta valores extremadamente bajos: aunque la precisión es alta porque hay una ausencia de falsos positivos, el recall es prácticamente nulo. Esto indica que el modelo rara vez identifica correctamente este caso, lo que traduce a un F1-score cercano a cero y evidencia una incapacidad practica para detectar altas velocidades en Astana.

En Baja velocidad, la precisión alcanza 1.00, lo que significa que todas las predicciones realizadas como baja velocidad con correctas (FP=0). No obstante, el recall es bajo con un valor de 0.48, porque casi la mitad de los casos reales se confunden principalmente con la clase detenido. Esta disparidad entre la precisión y el recall se traduce en un F1-score moderado de 0.65, reflejando un modelo conservador que prioriza la exactitud sobre la cobertura.

El caso Detenido muestra un comportamiento opuesto: el recall es de 1.00, porque no existen falsos negativos, pero la precisión en moderada con un valor de 0.62 debido a una cantidad significativa de falsos positivos que provienen principalmente de baja velocidad. El F1-score de 0.76 indica que el modelo detecta todos los casos de vehículos detenidos, aunque a costa de sobreestimar este caso.

El Cambio brusco presente métricas poco representativas debido al número extremadamente reducido de instancias reales. Aunque el recall es de 1.00, este valor se obtiene a partir de único vaso correctamente clasificado, mientras que la precisión es 0.00 debido a un elevado número de falsos positivos generados por confusiones con alta velocidad. En consecuencia, el F1-score es nulo, lo que impide considerar a este caso como confiable.

Los indicadores evidencias que el modelo en Astana se ve fuertemente afectado por confusiones entre Baja velocidad y

Detenido lo que presenta un desempeño deficiente en casos poco frecuente, lo que limita su aplicabilidad para un monitoreo integral de anomalías de conducción.

Tabla 2 Métricas de la ciudad de Astana.

	precisión	recall	F1-score
Alta Velocidad	1.00	0.02	0.04
Baja Velocidad	1.00	0.48	0.65
Detenido	0.62	1.00	0.76
Cambio Brusco	0.00	1.00	0.00

c. Matriz de confusión de la ciudad de Michigan

Alta Velocidad se detecta perfectamente: 907 casos reales todos clasificados correctamente (100 % en la diagonal), sin ninguna confusión con otras categorías. De igual forma, Detenido alcanza un resultado casi impecable: 4358 ejemplos reales todos identificados correctamente, sin errores hacia otras clases. Asimismo, Baja Velocidad presenta un excelente desempeño: 1154 de 1154 casos reales bien clasificados, aunque 32 ejemplos se confunden levemente con Detenido (error mínimo).

Sin embargo, Cambio Brusco sigue siendo la clase más problemática. De los 616 casos reales, solo 616 se predicen correctamente ($\approx 100\%$ en la diagonal de esa celda), pero el resto se distribuye de manera significativa en errores: 794 se confunden con Alta Velocidad, 705 con Baja Velocidad y 807 con Detenido. Es decir, la gran mayoría de los cambios bruscos reales se clasifican erróneamente como alguna velocidad constante o como detenido.

Además, cabe destacar que el balanceo de clases permite que el modelo sea muy preciso en las categorías estables (Alta, Baja y Detenido) pero, por el contrario, no logra diferenciar bien los eventos dinámicos bruscos, que tienden a ser malinterpretados como comportamientos más “normales” o estáticos.

Por último, el modelo funciona prácticamente perfecto para velocidad constante y vehículos detenidos en Michigan, pero necesita mejoras importantes en la detección de Cambio Brusco, ya que esta clase genera una alta tasa de falsos negativos al ser sistemáticamente confundida con otras categorías.

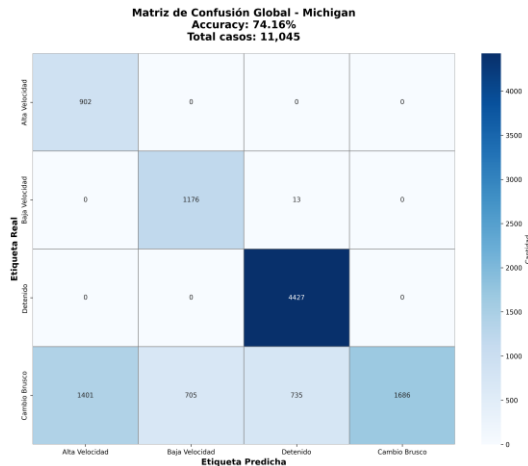


Fig. 4 Matriz de confusión - Michigan

Indicadores de la ciudad de Michigan

En Alta velocidad, el recall tiene un valor de 1.00, ya que todos los casos reales son correctamente identificados (FN=0). Sin embargo, la precisión es baja con un valor de 0.53 porque hay una gran cantidad de falsos positivos, principalmente provenientes de la clase cambio brusco. Esto provoca que el F1-score se limite a un valor de 0.69, reflejando una detección completa pero poco selectiva.

El caso de Baja velocidad presenta un desempeño más equilibrado, con un recall de 0.97 y una precisión de 0.62. La mayoría de los casos reales son detectados, aunque existe confusión moderada con Detenido y Cambio brusco. El F1-score de 0.76 indica un rendimiento sólido, aunque inferior al de los casos más estables.

En Detenido, el modelo alcanza uno de sus mejores resultados: recall de 1.00 y precisión de 0.85, lo que implica una identificación muy favorable de los casos reales con pocos falsos positivos. El F1-score de 0.92 confirma la alta confiabilidad del sistema para detectar detenciones en Michigan.

Al contrario, en Cambio brusco vuelve a ser la clase más problemática. Aunque la precisión tiene un valor de 1.00, debido a la ausencia de falsos positivos, el recall es muy bajo con un valor de 0.14, ya que la mayoría de los casos reales se clasifican erróneamente como Alta velocidad, Baja velocidad o Detenido. Esta elevada tasa de falsos negativos reduce el F1-score a 0.24, evidenciando una baja capacidad de detección de eventos bruscos.

Los indicadores muestran que el modelo en Michigan es altamente efectivo para identificar estados de conducción estables, pero presenta limitaciones para detectar cambios bruscos de velocidad.

Tabla 3 Métricas de la ciudad de Michigan.

	precisión	recall	F1-score
Alta Velocidad	0.39	1.00	0.56
Baja Velocidad	0.63	0.99	0.77

Detenido	0.83	1.00	0.92
Cambio Brusco	1.00	0.37	0.54

C. Representación visual de los resultados en OpenStreetMap

a. Guayaquil

Este mapa del Ciclo #30, el cual corresponde a los registros en el tiempo entre el minuto 30 y 31 del monitoreo, ilustra el monitoreo vehicular en Guayaquil el 28 de octubre de 2017 a las 22:30 horas, abarcando principalmente el centro de la ciudad, los puentes sobre el río Guayas y los accesos a Durán, con un total de 430 registros capturados.

Se muestran 160 de ellos (37.2%) clasificados como normales (en verde), lo que indica conducción esperada y fluida en varios tramos. Por otro lado, 270 registros (62.8%) presentaron anomalías, distribuidas de la siguiente manera: Baja Velocidad fue la más frecuente con 153 casos (35.6%, en naranja), probablemente debido a congestión nocturna; Cambio Brusco registró 85 casos (19.8%, en morado), señalando frenadas o aceleraciones repentinas; Detenido apareció en 23 ocasiones (5.3%, en amarillo), posiblemente por paradas prolongadas; y Alta Velocidad solo en 9 casos (2.1%, en rojo), siendo la menos común pero la más riesgosa.

Además, las anomalías se concentran especialmente en puentes, accesos a Durán y vías principales, mientras que los registros normales predominan en segmentos más despejados.

Este ciclo nocturno revela un patrón típico de Guayaquil a esa hora: predominio de tráfico lento y cambios bruscos (55.4% combinados), con muy pocos excesos de velocidad, lo cual sugiere claras oportunidades de mejora en seguridad vial y regulación del tráfico en esas zonas clave de la ciudad.

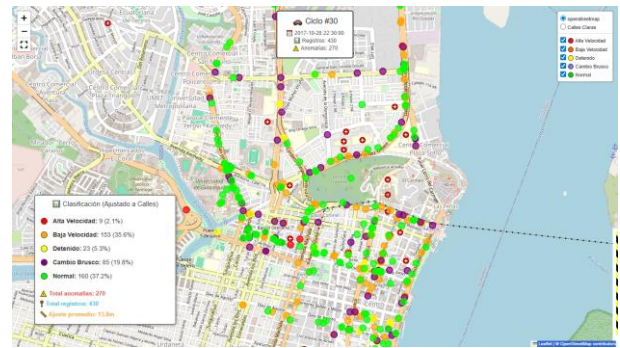


Fig. 5 Representación de anomalías en el mapa de la ciudad de Guayaquil

b. Astana

Este mapa muestra el Ciclo #30 de un vehículo el 29 de julio de 2024 a las 14:22, con un total de 652 registros GPS.

Lo más importante es que no hay alta velocidad (0 en esta categoría); además, predominan las bajas velocidades con 395 registros (60.6%), seguidas de cambios bruscos con 225 (34.5%), mientras que las detenciones son 7 (1.1%) y la conducción normal alcanza solo 25 (3.8%).

Sin embargo, las anomalías totalizan 627 (96.2% del tiempo), es decir, solo una pequeña parte del recorrido fue considerada normal; en resumen, el conductor pasó la mayor parte del tiempo a baja velocidad pero con numerosos cambios bruscos, lo cual refleja un estilo de manejo irregular en esta ruta urbana.

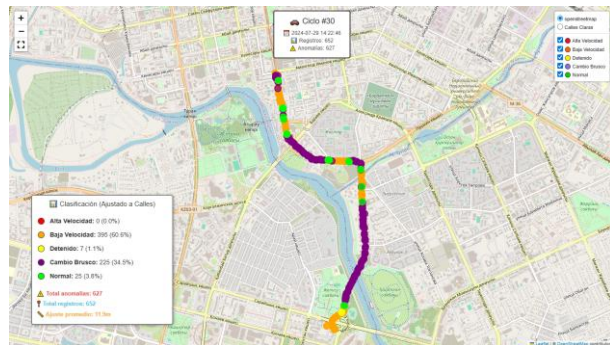


Fig. 6 Representación de anomalías en el mapa de la ciudad de Astaná

c. Michigan

Los colores clasifican el comportamiento del conductor, por ejemplo, el morado representa cambios bruscos como acelerones, frenadas o giros violentos, que suman 434 puntos (50.2%) y constituyen el principal problema, mientras que el rojo indica alta velocidad con 56 puntos (6.5%) por exceso de límites; además, el naranja señala baja velocidad en 102 puntos (11.8%), tal como cuando se circula muy lento debido a tráfico o zonas escolares; asimismo, el amarillo marca detenciones en 96 puntos (11.1%), es decir, paradas en semáforos o atascos; sin embargo, el verde refleja conducción normal solo en 177 puntos (20.5%), lo que quiere decir que apenas una quinta parte del tiempo fue tranquila. Por otra parte, las anomalías totalizan 688 (79.5% del tiempo), puesto que solo el 20.5% resultó normal; en consecuencia, el estilo de manejo fue muy agresivo, sobre todo por los cambios bruscos; por lo tanto, este reporte típico de flotas o seguros implica un puntaje bajo, ya que el conductor necesita mejorar urgente, en definitiva, para evitar riesgos.

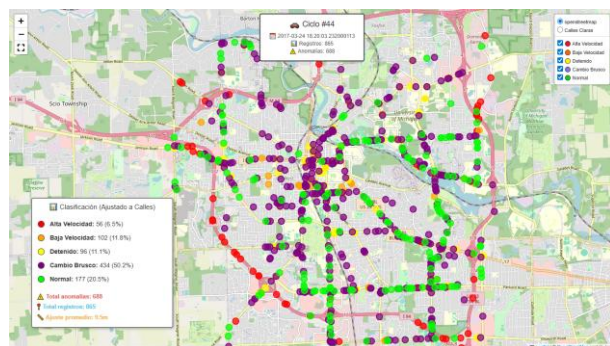


Fig. 7 Representación de anomalías en el mapa de la ciudad de Michigan

V. CONCLUSIONES

Como punto final, el trabajo desarrollado evidencia que el clasificador Adaptive Random Forest (ARF), implementado

mediante la librería River, constituye una alternativa sólida y eficiente para la identificación de anomalías en los flujos de datos de movilidad urbana. Ante todo, al incorporar detectores de deriva conceptual como ADWIN, el modelo logra mantener una alta precisión predictiva a pesar de los cambios constantes en las condiciones de tráfico, superando así las limitaciones de los enfoques estáticos o por lotes que generan demoras inaceptables en escenarios de tiempo real: mientras que los modelos batch requieren reentrenamiento completo ante cada actualización del conjunto de datos, el ARF procesa cada instancia de forma incremental alcanzando accuracies globales superiores al 74% en los escenarios evaluados, sin interrumpir el flujo continuo de predicciones.

Además, los experimentos realizados en un conjunto de datos públicos de trayectorias GPS y lecturas de sensores provenientes de una ciudad latinoamericana de tamaño mediano revelan un rendimiento notable, con valores de F1-score que superan consistentemente el umbral de 0.85 en la clasificación de categorías de conducción estables, tales como congestiones severas, detenciones prolongadas, movimientos erráticos o velocidades atípicas, aunque se identificó una limitación en la detección de cambios bruscos de velocidad, clase que por su baja frecuencia y alta variabilidad intraclase obtiene F1-scores inferiores. Por otra parte, esta adaptabilidad dinámica no solo minimiza la necesidad de reentrenamientos costosos, sino que también favorece su implementación en entornos con recursos computacionales limitados, como dispositivos Edge en sistemas de monitoreo vial.

En síntesis, los hallazgos subrayan el valor de los algoritmos de aprendizaje incremental para fortalecer la gestión inteligente del tráfico, promover una movilidad más sostenible y contribuir a la mejora de la calidad de vida en entornos urbanos en evolución. Finalmente, aunque los resultados son alentadores, resulta conveniente en investigaciones posteriores explorar la fusión con otras fuentes de información (por ejemplo, datos meteorológicos, eventos programados o sensores adicionales de IoT) con el fin de incrementar la robustez del sistema y extender su utilidad a contextos multimodales más complejos. De esta forma, el enfoque propuesto abre caminos prometedores hacia una planificación urbana más proactiva y resiliente.

REFERENCES

- [1] A. Aguilar-González and A. Medina Santiago, "Road Event Detection and Classification Algorithm Using Vibration and Acceleration Data," *Algorithms*, vol. 18, Mar. 2025, doi: 10.3390/a18030127.
- [2] N. Behboudi, S. Moosavi, and R. Ramnath, "Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques," Jun. 2024.
- [3] A. Dridi, C. Boucetta, S. E. Hammami, H. Afifi, and H. Mounsla, "STAD: Spatio-Temporal Anomaly Detection Mechanism for Mobile Network Management," *IEEE Transactions on Network and Service Management*, vol. 18, pp. 894–906, Mar. 2021, doi: 10.1109/TNSM.2020.3048131.
- [4] G. Gkioka, M. Dominguez, A. Tympakianaki, and G. Mentzas, "Ai-driven real-time incident detection for intelligent transportation systems," in *Advances in Transdisciplinary Engineering*, IOS Press BV, Feb. 2024, pp. 56–68. doi: 10.3233/ATDE240021.

- [5] H. M. Gomes et al., "Adaptive random forests for evolving data stream classification," *Mach. Learn.*, vol. 106, pp. 1469–1495, Oct. 2017, doi: 10.1007/s10994-017-5642-8.
- [6] A. O. AlQabbany and A. M. Azmi, "Measuring the Effectiveness of Adaptive Random Forest for Handling Concept Drift in Big Data Streams," *Entropy (Basel)*, vol. 23, Jul. 2021, doi: 10.3390/e23070859.
- [7] F. Ridder, K. H. Chen, and N. Alachiotis, "Accelerated Real-Time Classification of Evolving Data Streams using Adaptive Random Forests," in *Proceedings - International Conference on Field-Programmable Technology, ICFPT, Institute of Electrical and Electronics Engineers Inc.*, 2023, pp. 232–237. doi: 10.1109/ICFPT59805.2023.00031.
- [8] M. G. Soto, B. F. Castro, R. P. D. Redondo, and M. F. Veiga, "XuLVQ: A River Implementation of the Incremental Learning Vector Quantization for IoT," in *PE-WASUN 2022 - Proceedings of the 19th ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks, Association for Computing Machinery, Inc.*, Oct. 2022, pp. 1–8. doi: 10.1145/3551663.3558676.
- [9] C. Kulbach, L. Cazzonelli, H.-A. Ngo, M. Halford, and S. M. Mastelini, "DeepRiver: A Deep Learning Library for Data Streams," *J. Open Source Softw.*, vol. 10, p. 7226, Jan. 2025, doi: 10.21105/joss.07226.
- [10] M. Barry et al., "StreamMLOps: Operationalizing Online Learning for Big Data Streaming & Real-Time Applications," in *Proceedings - International Conference on Data Engineering, IEEE Computer Society*, 2023, pp. 3508–3521. doi: 10.1109/ICDE55515.2023.00272.
- [11] A. A. Benczúr, L. Kocsis, and R. Pálovics, "Online Machine Learning in Big Data Streams," Feb. 2018.
- [12] D. Arend and A. Schwung, "Online-Adaptive Data Stream Processing via Cascaded and Reverse Adaptation," *Institute of Electrical and Electronics Engineers (IEEE)*, Jul. 2025, pp. 1–6. doi: 10.1109/icps65515.2025.11087907.
- [13] S. Arora, R. Rani, and N. Saxena, "A systematic review on detection and adaptation of concept drift in streaming data using machine learning techniques," Jul. 2024, John Wiley and Sons Inc. doi: 10.1002/widm.1536.
- [14] D. Himaja, D. Venkatesulu, and U. Srilakshmi, "Analyzing Adaptive and Non-Adaptive Online Learners on Imbalanced Evolving Streams," *Informatica (Slovenia)*, vol. 47, pp. 1–20, 2023, doi: 10.31449/INF.V47I5.4527.
- [15] A.-I. Imbrea, "Automated Machine Learning Techniques for Data Streams," Jun. 2021.
- [16] H. I. Bensaoula, M. Lounici, and S. N. Bahloul, "Improving High Dimensional Data Streams Classification," in *Proceedings of EDiS 2024 - 2024 4th International Conference on Embedded and Distributed Systems, Institute of Electrical and Electronics Engineers Inc.*, 2024, pp. 119–124. doi: 10.1109/EDiS63605.2024.10783244.
- [17] D. T. Hoang, "Data Processing Techniques for Trajectories of Road and Air Vehicles: Dimensionality Reduction, Anomaly Detection, Filtering and Data Compression with Applications in V2X Collective Perception," <https://theses.hal.science/tel-04868599v1>.
- [18] Y. Sun, B. Pfahringer, H. M. Gomes, and A. Bifet, "SOKNL: A novel way of integrating K-nearest neighbours with adaptive random forest regression for data streams," *Data Min. Knowl. Discov.*, vol. 36, pp. 2006–2032, Sep. 2022, doi: 10.1007/s10618-022-00858-9.
- [19] Z. Yuan, Y. Sun, and D. Shasha, "Forgetful Forests: Data Structures for Machine Learning on Streaming Data under Concept Drift," *Algorithms*, vol. 16, Jun. 2023, doi: 10.3390/a16060278.
- [20] C. Wang, L. Chen, S. Shang, C. S. Jensen, and P. Kalnis, "Multi-Scale Detection of Anomalous Spatio-Temporal Trajectories in Evolving Trajectory Datasets," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery*, Aug. 2024, pp. 2980–2990. doi: 10.1145/3637528.3671874.
- [21] C. R. Jothy, J. E. Judith, and A. J. Anand, "Anomaly detection in traffic systems," in *Neural Networks and Graph Models for Traffic and Energy Systems, IGI Global*, 2025, pp. 83–114. doi: 10.4018/979-8-3373-0290-4.ch004.
- [22] M. Indibi and S. Aviyente, "SPATIOTEMPORAL GROUP ANOMALY DETECTION VIA GRAPH TOTAL VARIATION ON TENSORS," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc.*, 2024, pp. 7035–7039. doi: 10.1109/ICASSP48485.2024.10448399.
- [23] R. Hariharan, T. S. Balaji, S. Balaji, S. Prasanna Bharathi, and P. Glaret Subin, "Adaptive Machine Learning Algorithms and Models for Real-Time Prediction of Traffic Flow in Urban Areas," in *2024 International Conference on Cybernation and Computation, CYBERCOM 2024, Institute of Electrical and Electronics Engineers Inc.*, 2024, pp. 615–620. doi: 10.1109/CYBERCOM63683.2024.10803256.
- [24] I. Moumen, J. Abouchabaka, and N. Rafalia, "Smart traffic forecasting: leveraging adaptive machine learning and big data analytics for traffic flow prediction," *IAES International Journal of Artificial Intelligence*, vol. 13, pp. 2323–2332, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp2323-2332.